

## **Rapport projet Kaggle Chest X Ray image (Pneumonia)**

Projet préparé par

Maha GRAA

Crystal Renaud

Ensaf EL ERBI

# Sommaire

## Table des matières

I. Introduction	4
1) La pneumonie	4
2) Les signes de la maladie	4
a. Les signes cliniques	4
b. Les signes physiologiques	4
3) Le diagnostic	4
II. Matériels et Méthodes	6
1) Matériels	6
a. Les Datasets	6
b. Keras	6
c. Tensorflow	6
2) Méthodes	6
a. Téléchargement des Datasets à partir du website Kaggle	6
b. Preprocessing :	7
c. modèle CNN	7
d. Modèle ResNet50	7
III. Résultats et discussion	8
1) La visualisation des images	8
2) Visualisation de la distribution des DataSets	8
3) Visualisation des images Normalisées	10
4) CNN simple	10
5) ResNet	13
IV. Conclusion	14

# I. Introduction

## 1) La pneumonie

Selon l'organisation mondiale de la santé (OMS), la pneumonie représente 14 % de tous les décès d'enfants de moins de 5 ans, tuant 740 180 enfants en 2019. En effet, elle touche principalement les enfants de moins de 2 ans et les adultes de plus de 65 ans ayant un terrain prédisposant (asthme, tabagisme, BPCO (broncho-pneumonie chronique obstructive), déficit immunitaire ... ). La pneumonie est une infection respiratoire aiguë du tissu pulmonaire touchant les voies respiratoires basses. Elle peut être provoquée par des virus ou des bactéries et entraîner des complications comme des méningites.

La pneumonie est le plus souvent bactérienne et est causée par *Streptococcus pneumoniae*, le pneumocoque. Cette forme bactérienne touche régulièrement un seul des deux poumons. En revanche, la forme virale de la pneumonie touche les deux poumons de façon diffuse. La contamination peut se faire par inhalation de gouttelettes respiratoires contaminées ou bien par diffusion de la bactérie présente de façon commensale dans le rhino-pharynx vers les poumons.

## 2) Les signes de la maladie

### a. Les signes cliniques

Lors d'une pneumonie, qu'elle soit virale ou bactérienne, les symptômes sont de la toux, une gêne respiratoire avec ou sans fièvre, sensation de fatigue (asthénie)... Ils peuvent être similaires à ceux d'un rhume ou d'une grippe. Ces symptômes varient de légers à graves en fonction du type de germe, de l'origine de l'infection (communautaire ou nosocomiale), de l'âge du patient et de son état de santé avant l'infection. Dans les cas les plus graves, on peut observer une hyperthermie supérieure à 40°C, des douleurs thoraciques, une toux productive (expectorations purulentes voire hémoptoïques).

### b. Les signes physiologiques

Les symptômes de la pneumonie sont l'effet de l'inflammation du tissu pulmonaire. Lorsque la bactérie ou le virus pénètre l'organisme, celui-ci se défend grâce aux cellules immunitaires. Il y a alors une inflammation ainsi qu'une production de pus et de liquide qui empêche le passage normal de l'air dans les alvéoles, conduisant à la toux d'expectoration et aux difficultés respiratoires.

## 3) Le diagnostic

Avant de faire le diagnostic, le médecin interroge le patient sur ses antécédents médicaux (intervention médicale ou soin hospitalier récent, contact avec une personne malade ...), ainsi que sur ce qu'il ressent. Suite à cela, le diagnostic pourra se faire dans un premier temps grâce à une auscultation approfondie (recherche de bruit anormaux des poumons (râles)). Des examens complémentaires peuvent être demandés pour confirmer le diagnostic:

- Examen cyto bactériologique des crachats (ECBC): le recueil du crachat se fait le matin à jeun, après rinçage bucco-dentaire afin d'éviter la contamination par la flore buccale. Cet examen permet de déterminer la présence ou non de bactéries.
- Radiographie du thorax (chest X-ray): Cet examen permet une visualisation des tissus internes, des os et des organes, y compris les poumons. Lors d'une radiographie pulmonaire normale, les poumons apparaissent noir (ou très sombre) et les côtes sont blanches. Lors d'une infection, les poumons sont blanchis de façon un peu flou (coloration en verre dépoli). On pourra alors confirmer la pneumonie avec la présence de foyers infectieux.

- Analyse de sang (hémoculture): Cette analyse nous indiquera si l'infection est devenue systémique (passage de l'agent infectieux dans la circulation sanguine). Si l'infection est systémique, cela augmente la gravité de la maladie (risque de septicémie, de méningite ... ).
- Scanner thoracique: Cette procédure d'imagerie utilise une combinaison de rayons X et de technologie informatique pour produire des images horizontales, ou axiales, nettes et détaillées (souvent appelées tranches) du corps. Un scanner montre des images détaillées de n'importe quelle partie du corps, y compris les os, les muscles, la graisse et les organes. Les tomodensitogrammes sont plus détaillés que les radiographies ordinaires et permettent donc de mieux visualiser les foyers infectieux.

Un diagnostic bactériologique est demandé pour déterminer par la suite le traitement:

- Examen direct des prélèvements par ECBC ou par lavage broncho-alvéolaire avec un examen microscopique par connaître sa coloration GRAM et sa morphologie. Cette analyse permet de trouver le germe responsable.
- Mise en culture des prélèvements pour faire d'autres tests (culture aérobie ou anaérobie, en présence ou non de certains nutriments) et connaître la sensibilité aux traitements.

Le diagnostic de la maladie doit être rapide et précis afin d'éviter toutes les complications possibles. C'est dans l'optique de faciliter le diagnostic que ce projet a vu le jour. Son but est de pouvoir utiliser une Intelligence Artificielle avec le deep learning pour différencier les radiographies de poumon sain et de poumon atteint de pneumonie.

## II. Matériels et Méthodes

### 1) Matériels

#### a. Les Datasets

Des images radiographiques thoraciques ont été sélectionnées à partir des cohortes de patients pédiatriques âgés de un à cinq ans à partir du centre médical Guangzhou Women and Children's. Toutes les radiographies en question ont été réalisées dans le cadre des soins cliniques de routine des patients. Les radiographies pulmonaires de mauvaise qualité ou illisibles ont été supprimées.

Les diagnostics des images ont été évalués par deux médecins experts avant d'être autorisés à être exploités dans le système d'intelligence artificielle. Les radiographies sont également vérifiées par un troisième expert.

Le dataset contient 5863 images (JPEG) classées selon deux catégories (Pneumonia/Normal). Il est organisé en trois sous dossiers :

Train dataset : contient 1341 images classées "normale" et 3875 images classées "pneumonia"

Test dataset : contient 234 images classées "normale" et 390 images classées "pneumonia"

Validation dataset : contient 8 images classées "normale" et 8 images classées "pneumonia"

#### b. Keras

Keras est une bibliothèque qui a été développée dans le cadre du projet ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System).

Elle permet d'interagir avec les algorithmes d'apprentissage automatique et de réseaux de neurones profonds citons : Tensorflow , theano ...

Keras est conçu pour une expérimentation rapide sur les réseaux de neurones profonds grâce à son ergonomie et ses capacités d'extension.

#### c. Tensorflow

Tensorflow est un outil d'apprentissage automatique compatible avec le langage Python.

C'est une bibliothèque open source qui simplifie le processus d'acquisition des données, de génération, de prédictions et de raffinement des résultats futurs.

### 2) Méthodes

#### a. Téléchargement des Datasets à partir du website Kaggle

Le dataset en question est disponible sur le website Kaggle, directement sur la page d'accueil du Chest X-Ray Images (Pneumonia).

#### b. Exploration des Datasets

Après avoir téléchargé nos Datasets, l'exploration se fait en plusieurs étapes :

- affichage des exemples (normal , pneumonia) de chaque Dataset
- vérification de la distribution des données et affichage sous forme de plots.

### c. Preprocessing

Il n'était pas nécessaire d'améliorer la qualité des images car l'ensemble de celle-ci ont déjà une très haute résolution. Cependant, pour rendre le modèle plus robuste, une normalisation a été utilisée ainsi qu'une labellisation des images. On a attribué un label correspondant aux images "normales " et un label correspondant aux images " pneumonia".

On a commencé par définir un seed de 42 ce qui nous permettra d'assurer la reproductibilité des résultats à chaque compilation des modèles en question.

### d. modèle CNN

Nous présentons un modèle basé sur CNN affiné pour la classification de la pneumonie dans les images radiographiques thoraciques.

Les deux bibliothèques Python open source, Keras et TensorFlow, ont été utilisées pour entraîner le CNN en question dans le but de discriminer les caractéristiques pour la classification de la pneumonie à partir d'images radiographiques pulmonaires. Le même modèle CNN a été appliqué sur les images du Dataset traitées de deux façons différentes.

#### CNN sur images normalisées et reshapées :

Le modèle CNN appliqué est un modèle composé de quatres blocs convolutifs avec une activation non linéaire (ReLU), d'un Flatten afin d'aplatir nos images au sein de la couche Dense et d'une activation sigmoïde sur la dernière couche Dense.

On a appliqué ce modèle sur les images déjà normalisées et reshapées.

#### CNN sur images reshapées :

Le même modèle CNN décrit précédemment a été appliqué sur les images uniquement reshapées.

### e. Modèle ResNet50

Le ResNet-50 est une variante de l'architecture ResNet (residual neural network) composée de 50 couches profondes formées sur :

- au moins un million d'images de la base de données ImageNet.
- une initialisation aléatoire

Les sauts à double ou triple couche avec une activation non-linéaires (ReLU) et la normalisation par lots sont utilisés dans la plupart des modèles ResNet.

Au sein du modèle ResNet utilisé , l'application d'un Flatten permet d'aplatir les images .

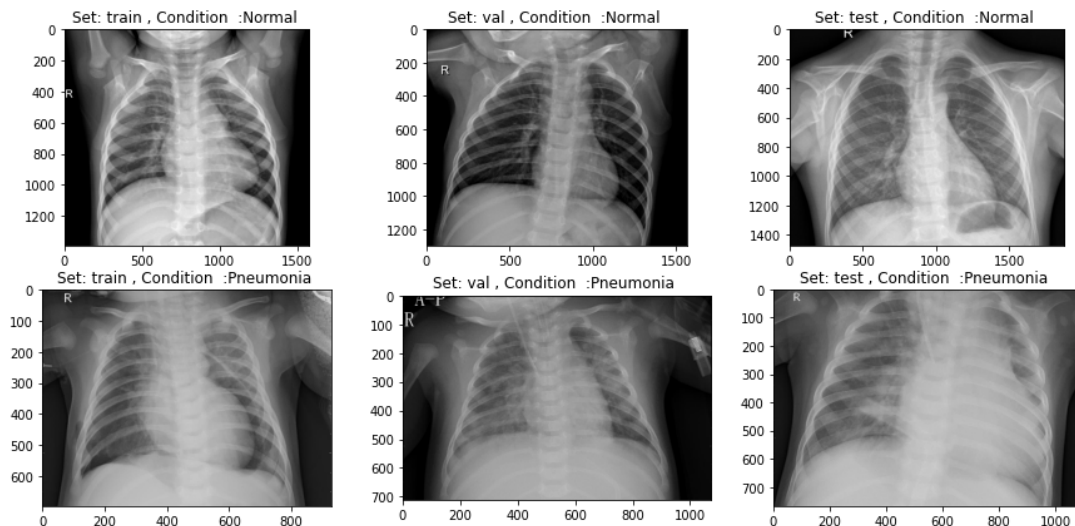
L'architecture du ResNet-50 appliqué consiste en des séquences de 3 blocs convolutifs avec une mise en commun moyenne.

### III. Résultats et discussion

#### 1) La visualisation des images

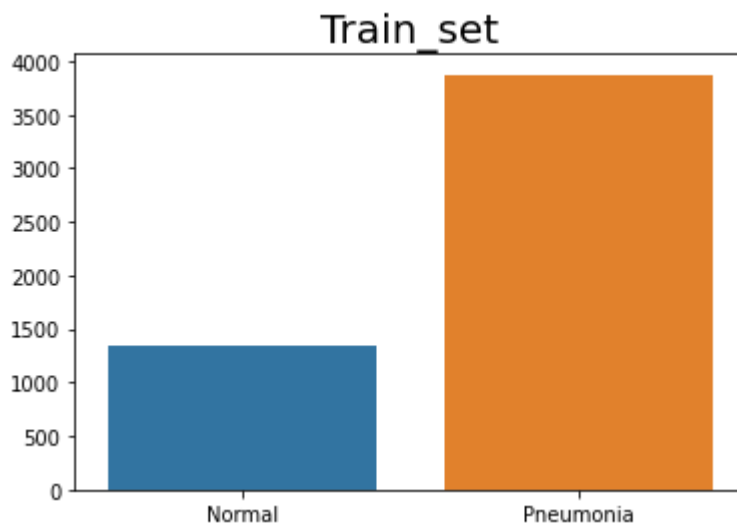
Pour chaque DataSet (train, validation et test), la visualisation deux exemples d'images est faite : Normal et pneumonia .

On constate que les images classées "pneumonia" sont plus opaques (aspect de verre dépoli aux niveaux des foyers infectieux donnant cette opacité au poumon).



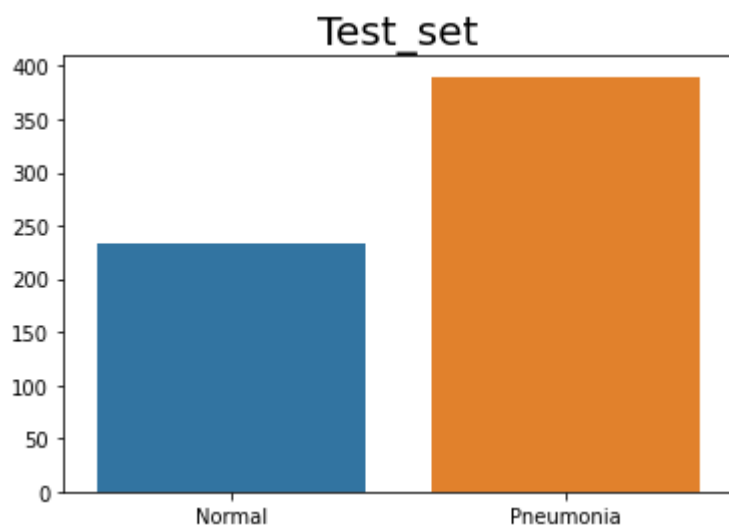
#### 2) Visualisation de la distribution des DataSets

a. Set : train , normal\_images : 1341 , pneumonia images : 3875



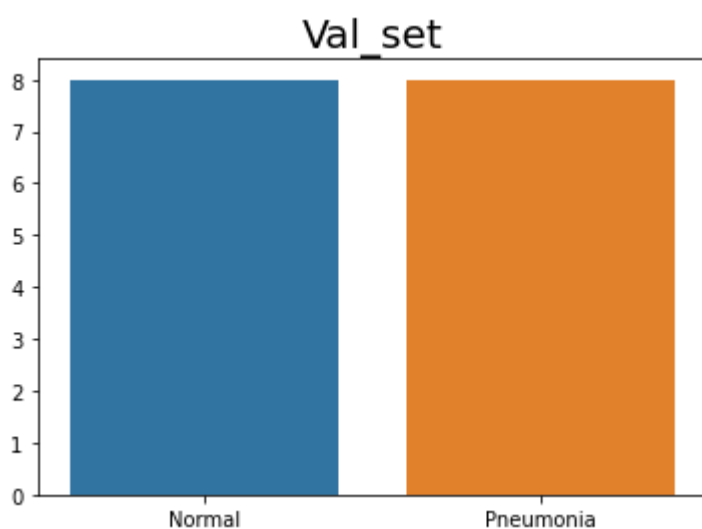
On constate une distribution inégale entre les images labellisées "Normal" et "Pneumonia" au niveau du Dataset Train. En effet, il y a trois fois plus d'image "Pneumonia" que d'image de poumon "Normal".

b. Set : test , normal images : 234 , pneumonia images : 390



On constate, là aussi, une distribution inégale entre les images labellisées “Normal” et “pneumonia” au niveau du Dataset test. Cette fois-ci, la quantité d’image “Normal” semble correspondre au deux tiers de celle “Pneumonia”.

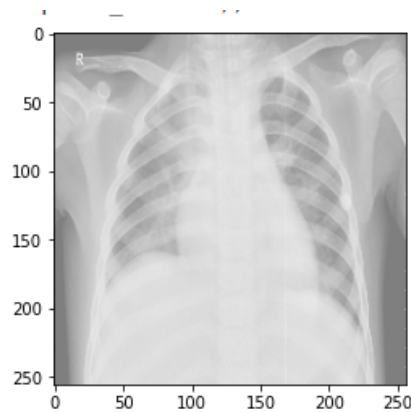
c. Set : val , normal images : 8 , pneumonia images : 8



On constate une distribution égale entre les images labellisées “Normal” et “pneumonia” au niveau du Dataset validation.



### 3) Visualisation des images Normalisées



Tout en sachant que la normalisation est un processus qui modifie la plage des valeurs d'intensité des pixels, on constate que l'image visualisée a perdu sa netteté du début. Il semble encore plus difficile pour un œil non entraîné de différencier les poumons sains des malades.

### 4) CNN simple

CNN sur images normalisées et reshapées :

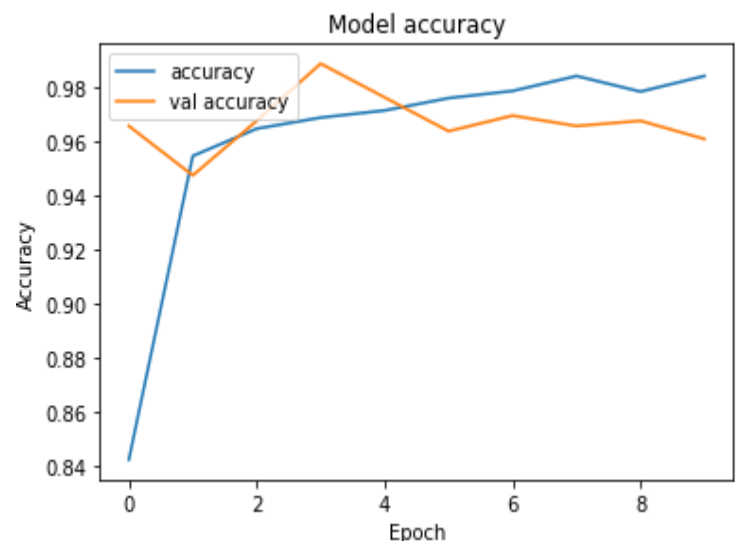
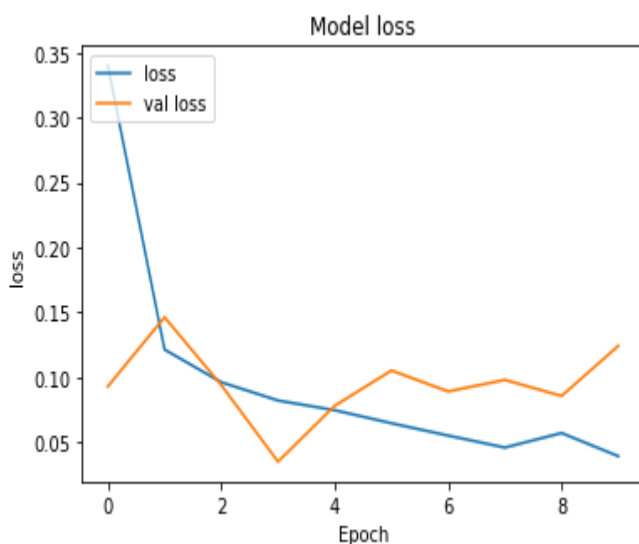
On a lancé le modèle CNN sur des images normalisées et reshapées en appliquant 10 epochs, un batch size de 32 et une validation split de 0.2 (on a appliqué une validation split de 0.2 parce que notre dataset validation n'est pas suffisante, ce qui conduisait à une mauvaise validation de l'apprentissage de nos neurones).

D'après le modèle "loss" obtenu on constate que la loss diminue au cours des epochs, elle est passée de 0.34 au début et est allée jusqu'à 0.038 à la dernière epoch.

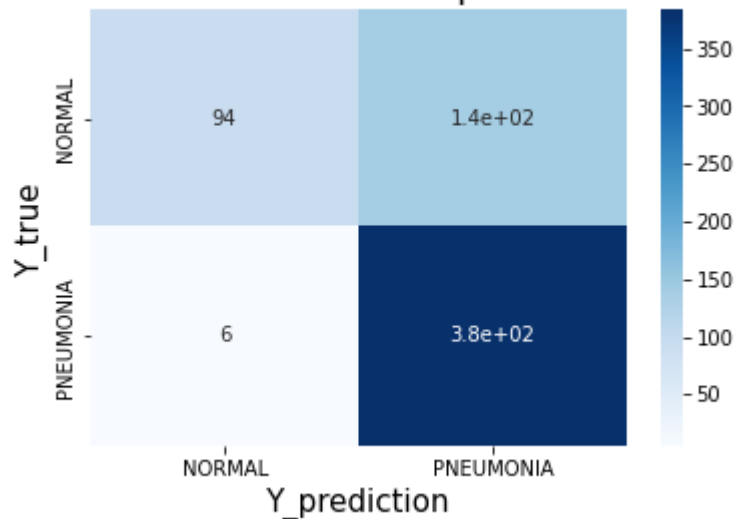
La validation loss nous montre qu'au niveau des dernières epochs, il existe un phénomène d'overfitting qui est peut être dû à un nombre un peu élevé d'epochs qui mène au surapprentissage.

D'après le modèle "accuracy" obtenu on constate que l'accuracy augmente au cours des epochs, elle passe de 0.84 à 0.98.

La validation accuracy reste assez stable au cours des epochs (environ 0.96). On constate tout de même un phénomène d'overfitting au niveau des dernières epochs.



Confusion matrix of CNN reshaped and normalized



La matrice de confusion montre que sur 234 images ayant le label “Normal “, le modèle est capable de détecter seulement 94 images normales , ce qui donne une spécificité de 40% (spécificité = vrai négatif / poumon sain). Ce qui signifie que plus d’une fois sur deux, le modèle dit qu’un poumon sain est atteint de pneumonie.

Et parmi 390 images ayant le label “pneumonia”, le modèle est capable de détecter 384 images “pneumonia”, ce qui représente une sensibilité de 98.44% (sensibilité = vrai positif / poumon malade).

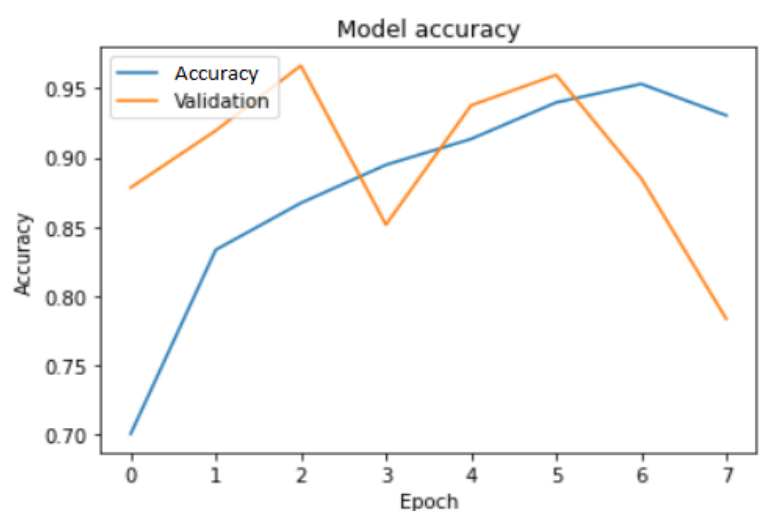
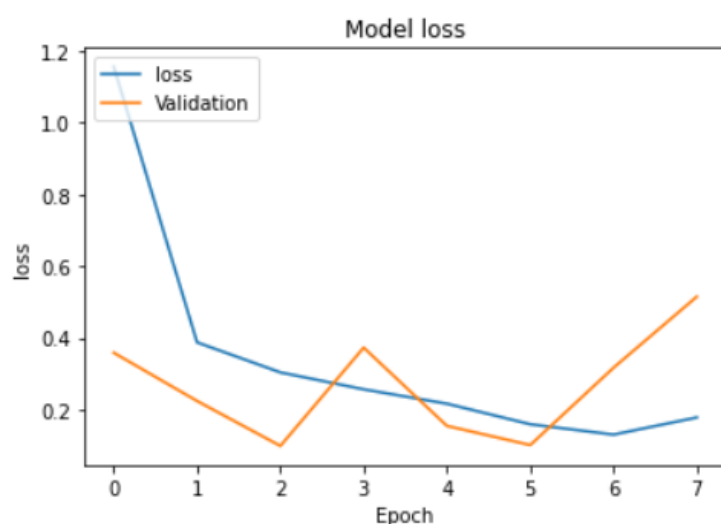
```
model.evaluate(X_test_norm_reshape,y_test_norm_reshape)
```

```
20/20 [=====] - 7s 327ms/step - loss: 1.3554 - accuracy: 0.7660  
[1.3554086685180664, 0.7660256624221802]
```

En évaluant le modèle , on a obtenu un loss de 1.3554 et une accuracy de 0.766. Ces résultats montrent que l’apprentissage du réseau de neurones n’est pas optimal.

### CNN sur images reshapées :

Le modèle CNN est lancé une seconde fois sur des images reshapées en appliquant 8 epochs, un batch size de 32 et une validation split de 0.2.



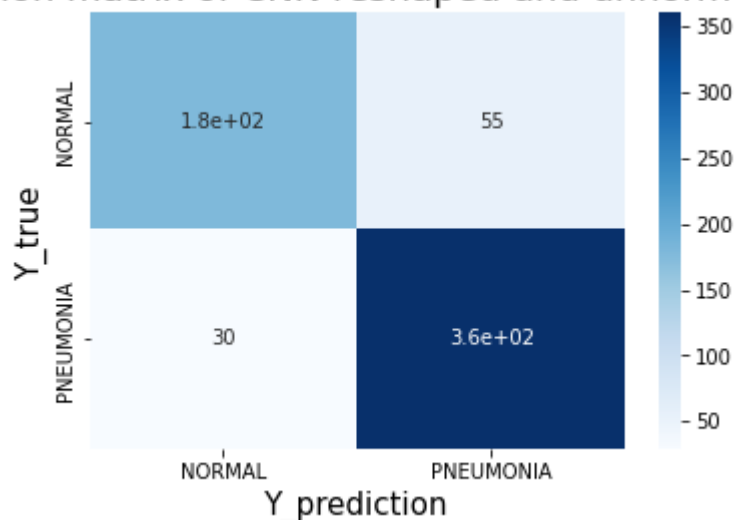
D'après le modèle “loss” obtenu on constate que la loss diminue avec les epochs, il est passé de 1.15 au début à 0.17 à la dernière epoch.

La validation loss nous montre qu’au niveaux des dernières epochs on a un phénomène d'overfitting qui peut être dû un nombre un peu élevé d’epochs qui mène à un surapprentissage.

D'après le models “accuracy” obtenu on constate que l’accuracy augmente au cours des epochs.

La validation accuracy semble rester stable sur les premières epochs avant de diminuer, on constate un phénomène de overfitting aux niveau des dernières epochs.

Confusion matrix of CNN reshaped and unnormalized



La matrice de confusion montre que sur 234 images ayant le label “Normal “, le modèle est capable de détecter 179 images normales, ce qui représente une spécificité de 76.4%.

Et parmi 390 images ayant le label “pneumonia”, le modèle est capable de détecter 360 images “pneumonia”, ce qui représente une sensibilité de 92 %.

```
model.evaluate(x_test_reshape, y_test_reshape)
```

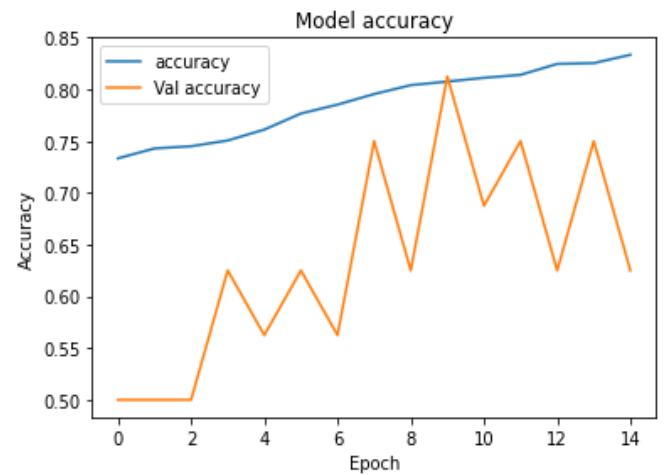
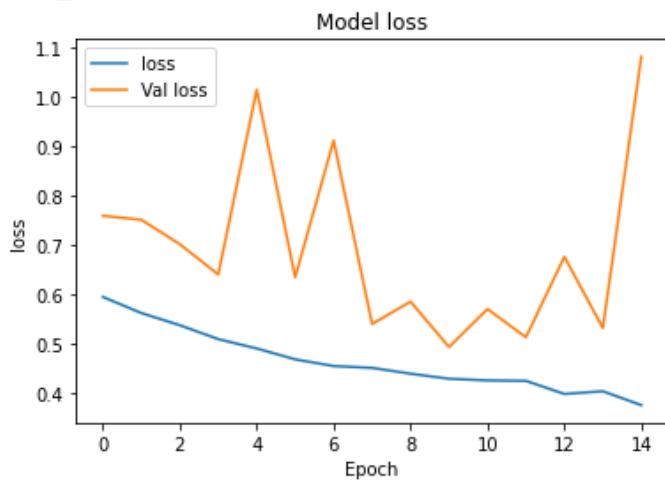
```
20/20 [=====] - 2s 100ms/step - loss: 0.3196 - accuracy: 0.8574  
[0.31955140829086304, 0.8573718070983887]
```

En évaluant le modèle , on a obtenu un loss de 0.319 et une accuracy de 0.857.

En comparant l’attribution des images reshapées au modèle avec l’attribution des images reshapées et normalisées, on constate que les résultats sont meilleurs sur les images non normalisées. Ceci s’explique par le besoin d’avoir une netteté optimale pour différencier les poumons normales des poumons pneumonia.

## 5) ResNet

Dans le but d’évaluer notre modèle de ResNet, la loss et l’accuracy ont été déterminées. Les résultats obtenus sont illustrés par les figures ci-dessous.



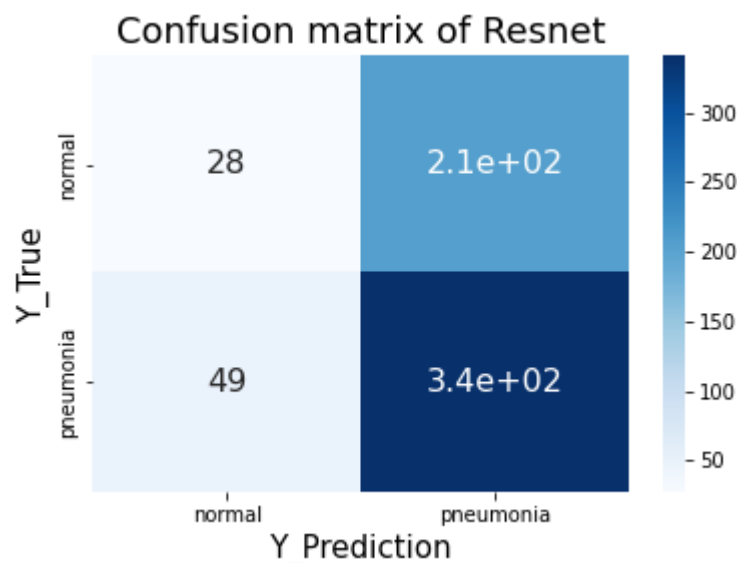
Le modèle ResNet est lancé sur des images normalisées et reshapées en appliquant 15 epochs, un batch size de 16. Les images sont préalablement traitées par un générateur qui les normalise et les redimensionne. Ce générateur permet de réduire la consommation de mémoire, de paralyser les processus pour les accélérer et d'avoir une meilleure cohérence et lisibilité du code.

En observant le modèle “loss”, on constate qu’il diminue au cours des epochs. Il est passé de 0,6 à une valeur inférieure à 0,4.

Ce qu’on peut noter aussi c’est qu’on a une loss de validation qui est beaucoup plus grande que la loss, ce qui rend le graphe ininterprétable. Cette interprétabilité est peut-être due à une mauvaise capacité du modèle à apprendre en profondeur sur l'ensemble de validation. Surtout que, pour le Resnet, le dataset de validation fourni par kaggle, et composé de seulement 16 images, est gardé.

L’accuracy obtenue pour ce modèle a augmenté tout au long des epochs, elle est passée de 0,73 à une valeur proche de 0,85.

La validation accuracy permet d’obtenir une mesure de précision sur l’ensemble de validation. On constate qu’elle augmente de façon anarchique au cours des epochs avec des chutes rapide suivi de remontées.



La matrice de confusion du ResNet montre que sur 234 images normales, le modèle a réussi à identifier correctement seulement 28 images, ce qui représente une spécificité de 11,96%.

Parmi les 390 images de pneumonie, le modèle de ResNet a pu reconnaître 341 images “pneumonia” ce qui représente une sensibilité de 87,4% .

Après qu’il ait tourné pendant 2 heures et 54 minutes, le modèle ResNet, appliqué sur les données du test, nous a donné une loss de l’ordre de 0,56 et une accuracy de 0,7388.

```
model2.evaluate(test_generator2)
```

```
5/5 [=====] - 122s 26s/step - loss: 0.5596 - accuracy: 0.7388 - precision: 0.7075 - recall: 0.9923  
[0.5596155524253845,  
0.7387820482254028,  
0.7074954509735107,  
0.9923076629638672]
```

## IV. Conclusion

Au cours de ce projet, nous nous sommes intéressés au sujet riche et passionnant de l’imagerie médicale. L’application du Deep learning à des problèmes d’actualité comme la détection du Covid19 ou la détection du cancer des poumons peut être importante.

On a réussi à faire deux modèles de classification dans le but de différencier les radiographies des poumons sains et des poumons atteints de pneumonie :

- Un CNN simple qui a été appliqué dans un premier temps sur les images normalisées et reshapées, et dans un deuxième temps sur les images seulement reshapées. Les résultats obtenus ont montré un meilleur rendement, en termes de Loss et d’Accuracy, avec les images reshapées qu’avec celles normalisées. Ceci est sûrement dû à la perte de netteté de l’image suite à la normalisation. En effet, l’apport de bruit sur l’image empêche d’apprendre correctement pour le réseau.
- Un ResNet appliqué sur des images normalisées et reshapées par l’intermédiaire d’un générateur. Les résultats obtenus montrent que le modèle semble être incapable de détecter les images normales. En effet, on a obtenu un pourcentage de 88,03% de faux positifs, ce qui montre clairement que le réseau ne reconnaît pas les radiographies de poumon sain.

Nous avons réussi à créer des réseaux qui classifient les radiographies de poumon en sain et pneumonie à partir d’une collection d’images radiologiques en utilisant un modèle CNN et un modèle ResNet. D’après les résultats obtenus, on peut dire que le modèle CNN appliqué sur les images reshapées est plus performant que le ResNet. Par la suite, on pourrait envisager de faire le Resnet avec plus d’époque et sur des images non normalisées pour que celle-ci restent nettes afin d’avoir des résultats satisfaisants. On pourrait aussi voire à augmenter les données du Dataset de validation afin de diminuer le phénomène d’overfitting et de mieux évaluer les performances du modèle.

## V. Bibliographie

Site ameli sur la pneumonie:

<https://www.ameli.fr/assure/sante/themes/pneumonie/definition-facteurs-risque>

Site de l'OMS sur la pneumonie:

<https://www.who.int/fr/news-room/fact-sheets/detail/pneumonia#:~:text=La%20pneumonie%20se%20d%C3%A9fini%20comme,personne%20en%20bonne%20sant%C3%A9%20respire.>

Site de Keras:

<https://keras.io/>