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.

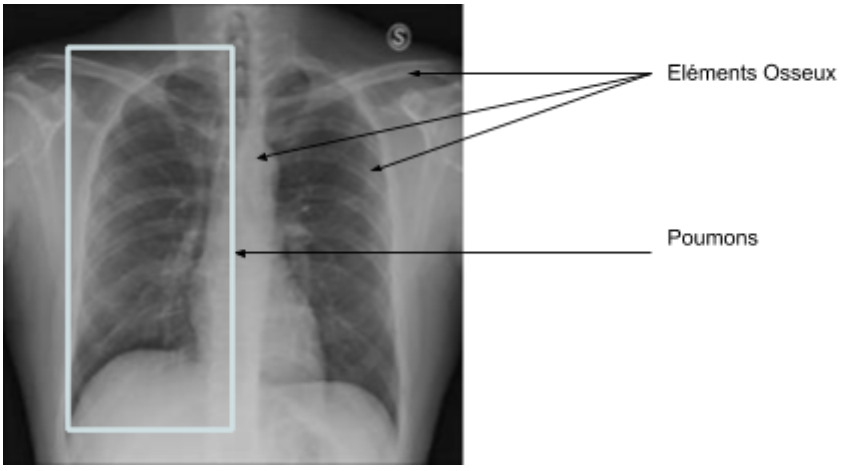
L'analyse exploratoire se décompose en 3 étapes à savoir l'analyse brute du dataset ; l'analyse métrique du dataset et une première identification des solutions pouvant être envisagées afin de pallier au manquement du dataset.

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

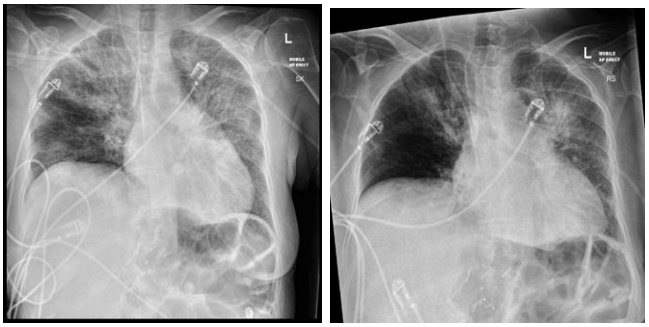
décline au niveau du contenu, de la résolution des images et du contraste/luminosité.

Si l'on s'intéresse au contenu, les images sont des radiologies 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.



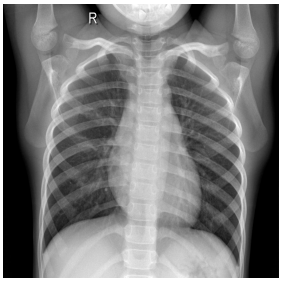
Au-delà de ces éléments, certains clichés contiennent des informations supplémentaires (annotations) ou témoignent de la présence de dispositifs médicaux lors de la prise des clichés.

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

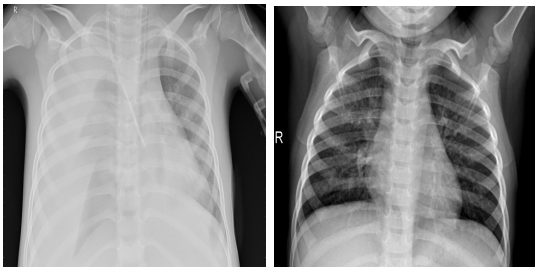


Concernant la **qualité et la résolution**, nous pouvons observer qu'il existe une disparité flagrante inter-classes et intra-classes notamment concernant la classe "Covid19".

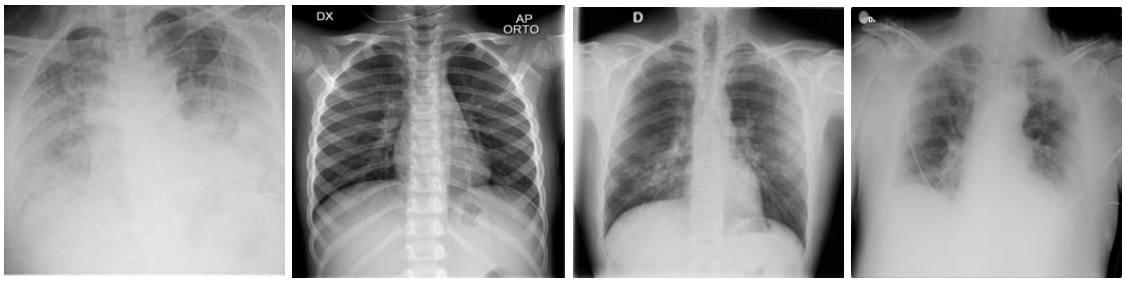
Ainsi, les clichés de la classe 'Normal' se caractérisent par un très grande homogénéité en termes de contrastes et de résolution. Cette dernière est élevée, les éléments distinctifs des clichés (poumons, éléments osseux) sont facilement identifiables.



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.

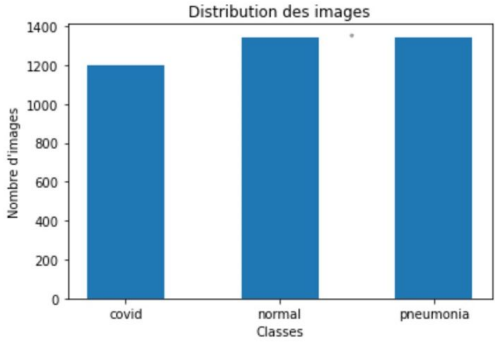
2) Analyse quantitative des données

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.

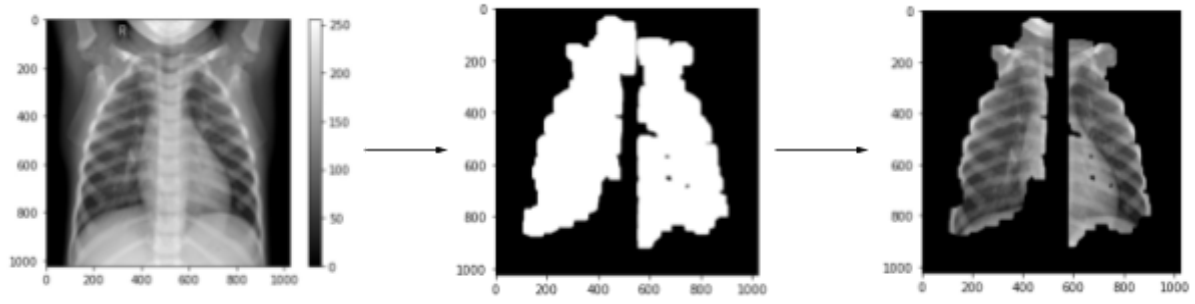
```
Entrée [14]: df['luminosite'].describe()

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 ?