Le dataset provient du **COVID-19 Radiography Database** et contient des radiographies pulmonaires réparties en 4 classes :

* **COVID :** patients infectés par le COVID-19 🡸 **Variable Cible du modèle**
* **Normal :** patients sains
* **Lung\_Opacity :** patients présentant une autre pathologie aux poumons
* **Viral Pneumonia :** patients infectés par une pneumonie virale (autre que le COVID-19)

Les images sont toutes au format .png avec une taille de 299x299 pixels.

Chaque image dispose d’un masque associé (délimitant les poumons), au format .png avec une taille de 256x256 pixels.

Les métadonnées associées sont fournies dans des fichiers .xlsx :

* Nom du fichier (Identifiant unique)
* Taille de l’image
* Source (URL)
* Format de l’image

**Répartition des classes**

Le dataset présente un **fort déséquilibre** des classes :

* La classe **Normal** est largement sur-représentée (~10 000 images).
* **COVID** et **Viral Pneumonia** sont sous-représentées.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

**Implication** : Il sera nécessaire de compenser ce déséquilibre par des techniques comme le suréchantillonnage ou la pondération des classes lors de la modélisation. Il faudra aussi utiliser une métrique d’évaluation adaptée (comme le score F1) et des séparations de données (entrainement/validation) stratifiées pour conserver l’équilibre des classes.

**Provenance des images :**

Une image contenant texte, capture d’écran, logiciel, conception

Le contenu généré par l’IA peut être incorrect.

Le graphique ci-dessus met en évidence plusieurs constats :

* Les classes **Viral Pneumonia** et **Lung Opacity** n’ont qu’une source unique (différente)
* La classe **Normal** se répartie sur les deux sources précédentes.
* La classe **COVID** se répartie sur 6 autres sources (toutes différentes, dont une prépondérante)

Cela est susceptible d’impliquer des biais d’entrainement du modèle.

**Exemples d’images**

Une image contenant film radiographique, Imagerie médicale, radiologie, radiographie

Le contenu généré par l’IA peut être incorrect.

Les visualisations montrent une **variabilité importante** dans la qualité des images :

* Certaines images sont bien contrastées et bien alignées.
* D'autres sont floues, sous-exposées, ou mal alignées.

**Implication** : Un prétraitement des images peut-être nécessaire pour assurer une meilleure cohérence et homogénéité des données avant leur entrée dans les modèles.

**Luminosité**

**Une image contenant capture d’écran

Le contenu généré par l’IA peut être incorrect.**

Le graphique ci-dessus représente la répartition des images en fonction de leur luminosité (intensité moyenne). Les classes **Viral Pneumonia** et **COVID** présentent une répartition quasi-normale, alors que les classes **Normal** et **Lung Opacity** présentent une 2nd pic d’intensité (bimodal) plus élevée. Le tracé en fonction de la provenance des données ne met pas en évidence une influence de celle-ci. (Pour **Lung\_Opacity** : source unique, pour **Normal** : graphique ci-dessous) :

Une image contenant capture d’écran, texte, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

**Contraste global par classe**

Une image contenant diagramme, capture d’écran, Rectangle, texte

Le contenu généré par l’IA peut être incorrect.**Une image contenant capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.**

La répartition du contraste est normale pour les classes **Viral Pneumonia**, **COVID**, et **Lung Opacity** (avec des écart-types assez élevés sur ces deux dernières) et présentent une moyenne à peu près identique. En revanche, pour la classe **Normal**, les images sont d’avantage contrastée, et la répartition s’écarte d’une loi normale

**Implication** : Le contraste pourrait être une variable discriminante utile pour la classification.

**Analyse de redondance**

Une **analyse de redondance** (par hash perceptuel) a détecté **132 groupes de doublons** :

* Principalement dans la classe **COVID**.
* Quelques cas dans **Normal** et **Viral Pneumonia**.

Exemples :

|  |  |  |
| --- | --- | --- |
| Une image contenant film radiographique, Imagerie médicale, radiologie, Rayon X  Le contenu généré par l’IA peut être incorrect.  1 : COVID-1.png | Une image contenant film radiographique, Imagerie médicale, radiologie, Rayon X  Le contenu généré par l’IA peut être incorrect.  2 : COVID-140.png | Une image contenant film radiographique, Imagerie médicale, radiologie, Rayon X  Le contenu généré par l’IA peut être incorrect.  3 : COVID-591.png |

**Implication** :

* Risques de fausses performances si des doublons se retrouvent dans les ensembles d’entraînement et de test.
* Surreprésentation de certains cas cliniques.

**Actions recommandées** :

* Marquer ou supprimer les doublons.
* Répartir les doublons de manière contrôlée pour éviter la fuite d'information.

**Histogramme de la surface de masque**

Les tailles de masques (~surface pulmonaire) sont correctement équilibrées selon une loi normale pour l’ensemble des classes, avec une moyenne sensiblement plus élevée pour la classe **Normal** (dont la significativité n’est pas démontrée)

Une image contenant diagramme, Tracé, capture d’écran

Le contenu généré par l’IA peut être incorrect.

**Exemple utilisation masque**

En s’intéressant uniquement à la zone d’intérêt délimitée par le masque selon l’exemple suivant :

Une image contenant film radiographique, noir, capture d’écran, noir et blanc

Le contenu généré par l’IA peut être incorrect.

Il est possible de recalculer les valeurs de luminosité et contraste ‘pulmonaires’, dont voici les résultats :

Une image contenant texte, Tracé, capture d’écran, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

La luminosité ne semble pas trop affectée par le masque (on retrouve à peu près la même distribution). En revanche pour le contraste, la répartition n’est plus normale sur la classe COVID et sur la classe Normal.

**Test statistique réalisé :**

Pour chaque propriété, un test ANOVA a été réalisé pour déterminer la significativité statistique de l’appartenance à l’une des classes sur la propriété. Les résultats montrent une p-value << 0.05, ce qui confirme l’effet significatif des classes sur ces propriétés.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

**Synthèse des points clés**

| **Aspect analysé** | **Observation principale** | **Action recommandée** |
| --- | --- | --- |
| **Répartition des classes** | Déséquilibre fort (Normal > 2x COVID) | Pondération / data augmentation |
| **Luminosité** | Images peu lumineuses dans certaines classes | Ajustement des luminosités |
| **Contraste** | Variabilité plus élevée dans COVID | Utiliser comme feature pour modélisation |
| **Redondance** | Doublons intra-classe (principalement COVID) | Suppression / répartition attentive |

**Conclusion**

Le dataset est globalement propre : il n’y a pas de données manquantes, juste quelques doublons à gérer. Les propriétés des images (luminosité, contraste) et des images masquées ont été caractérisées, il y’a des disparités statistiquement significatives suivant les classes.