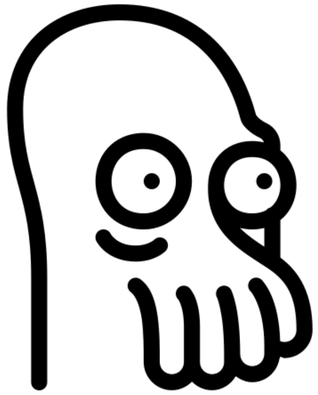


Zoidberg 2.0

Projet Intelligence Artificielle et Big Data

EPITECH.
L'ECOLE DE L'EXPERTISE INFORMATIQUE

Introduction

Notre objectif sur ce projet

Notre but est de développer une solution de machine learning fiable et robuste afin d’assister les médecins dans le dépistage de la pneumonie chez les patients.

Introduction

Le plan de notre présentation

Partie #01

Présentation du dataset et exploration des données.

Partie #02

Présentation des différents model de machine learning explorés.

Partie #03

Analyse des résultats et comparaison des différentes méthodes.

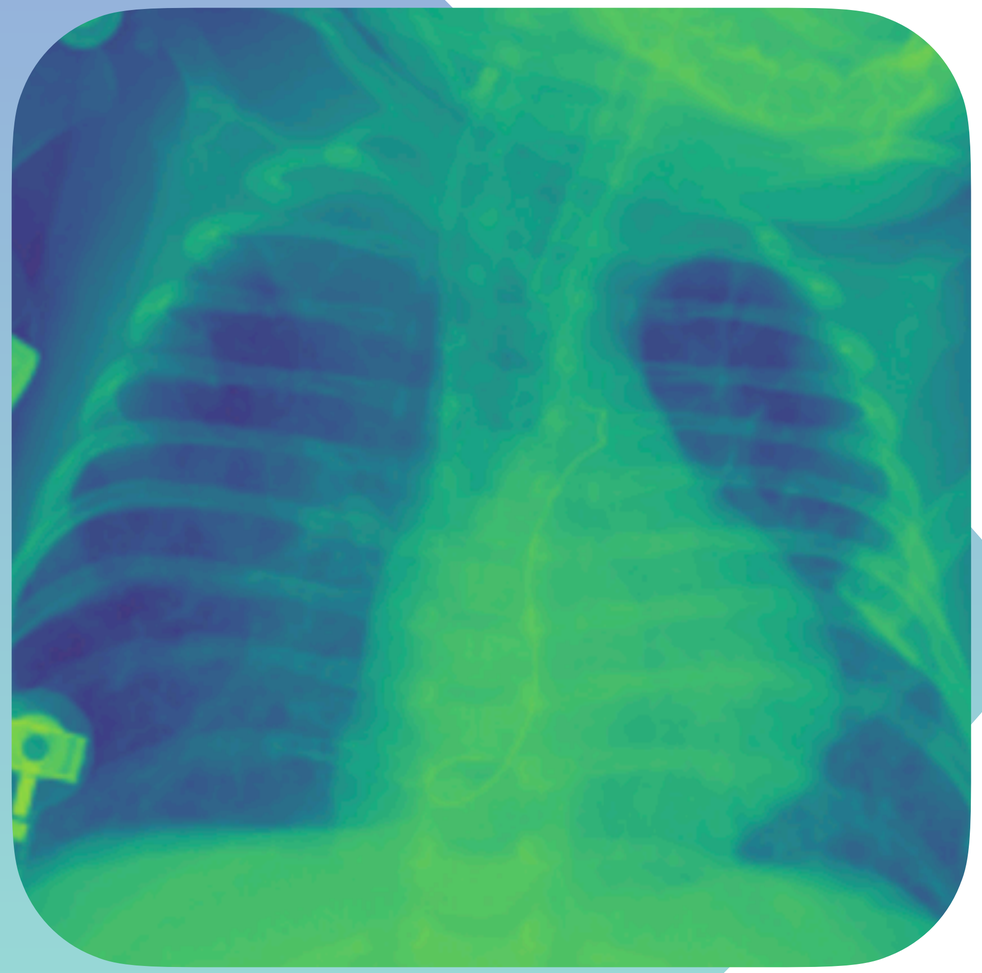
Exploration des données

Dataset et traitement des données

Dataset

Un ensemble de données classées par catégories selon certaines caractéristiques prédéfinies.

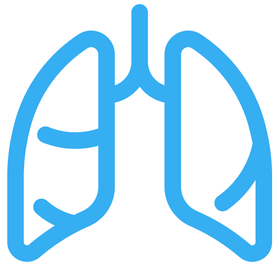
En machine learning, cela permet d’entrainer les modèles.



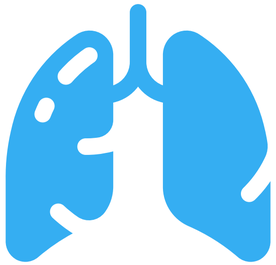
EXPLORATION DES DONNÉES

Présentation

Chaque dataset représente un ensemble d’images de radio des poumons:



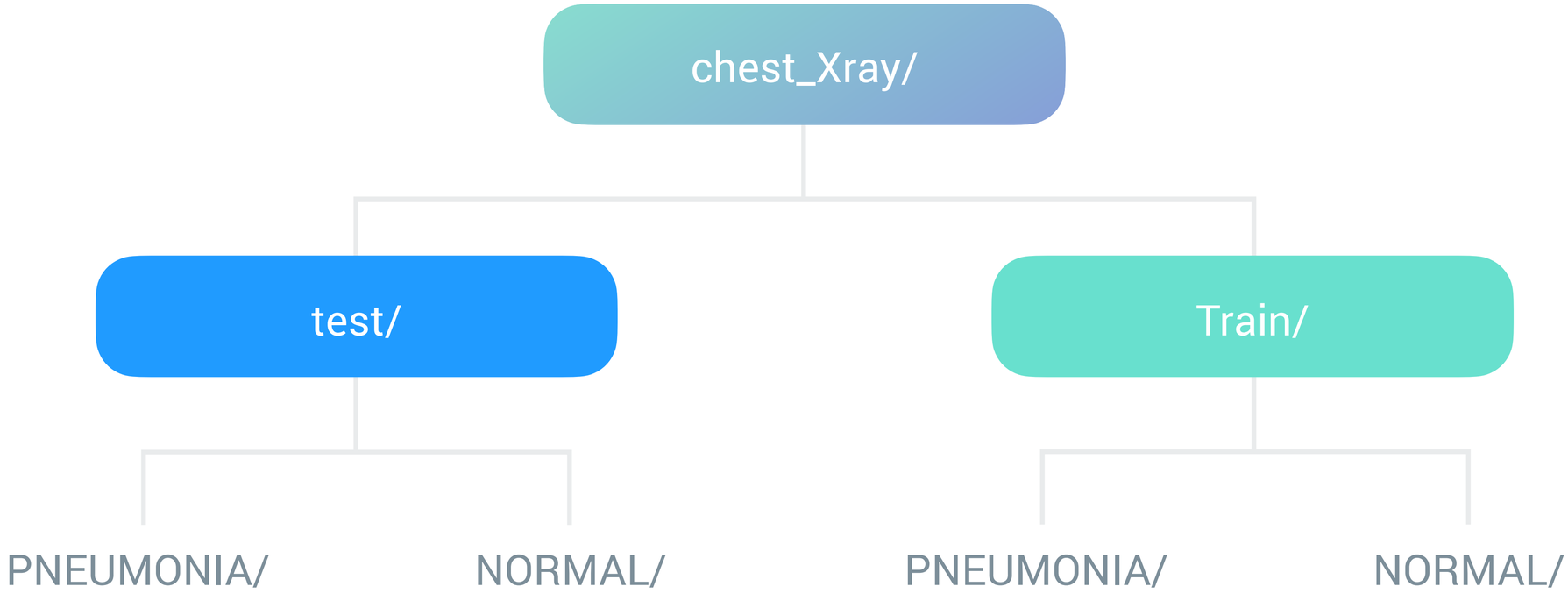
Non infectés (sains)



Infectés (virus ou bactérie)

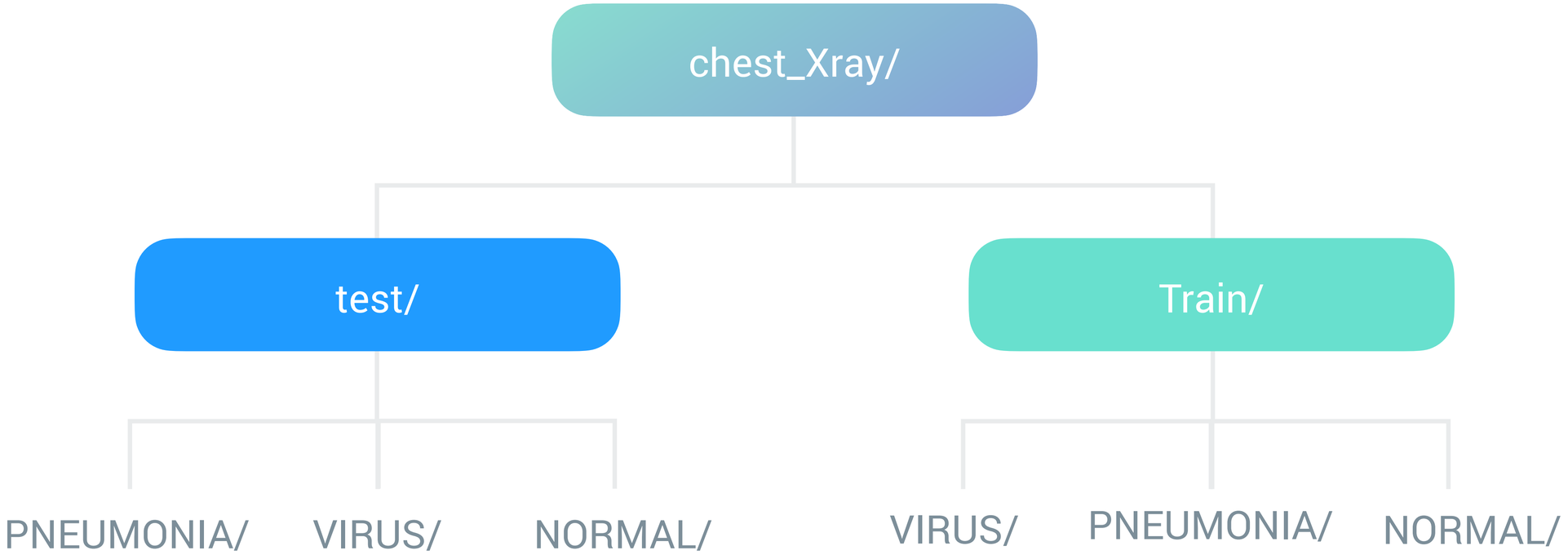
EXPLORATION DES DONNÉES

Structure du dataset



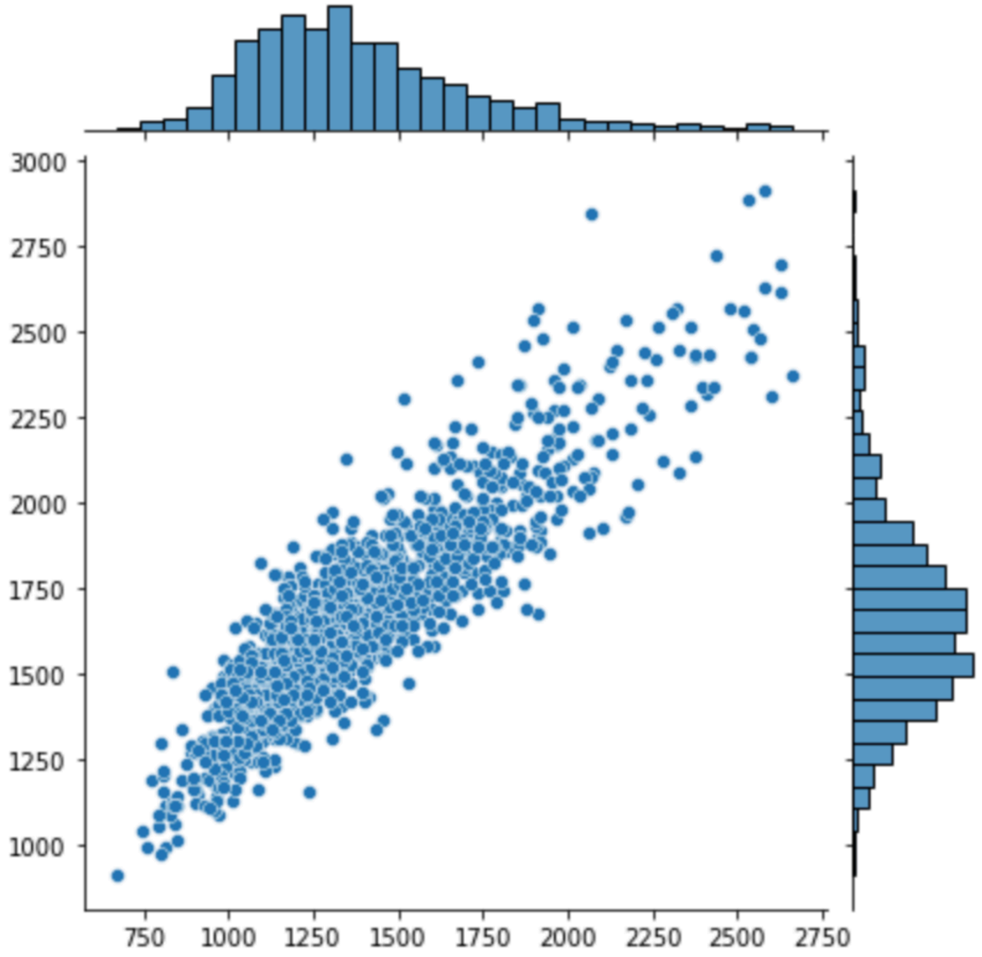
EXPLORATION DES DONNÉES

Nouvelle structure du dataset



EXPLORATION DES DONNÉES

Visualisation du format des images



- Trop grande disparité dans les dimensions des images

- Nos models ont besoin d'être entrainé sur des images aux dimensions identiques

EXPLORATION DES DONNÉES

Traitement des données

Mise à l’échelle

C’est manipuler les images dans le but d’obtenir les mêmes dimensions

Normalisation

Faire tenir les valeurs de chaque pixel entre 0 et 1

EXPLORATION DES DONNÉES

Traitement des données

Diminution des dimensions

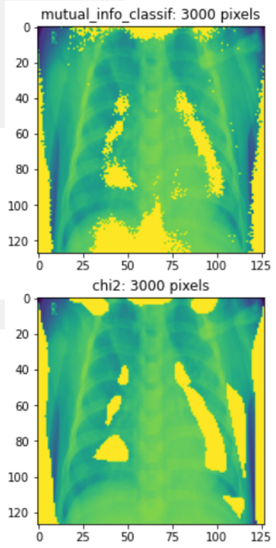
Supprimer les pixels qui n’apportent pas d’informations

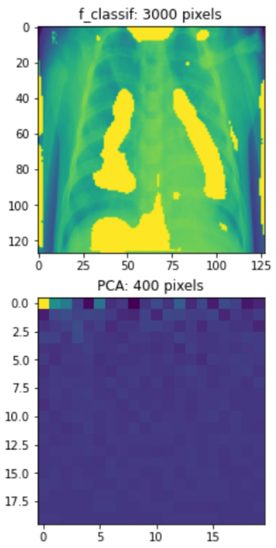
Augmentation des données

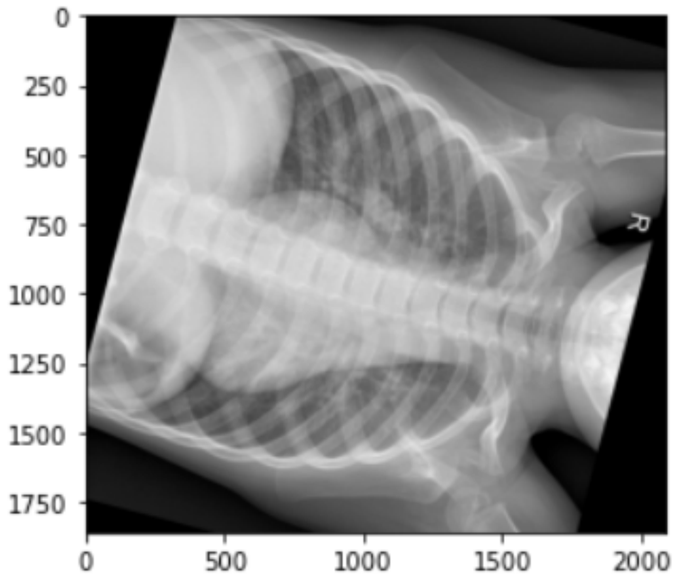
Permet de prévenir le sur-apprentissage (over-fitting)

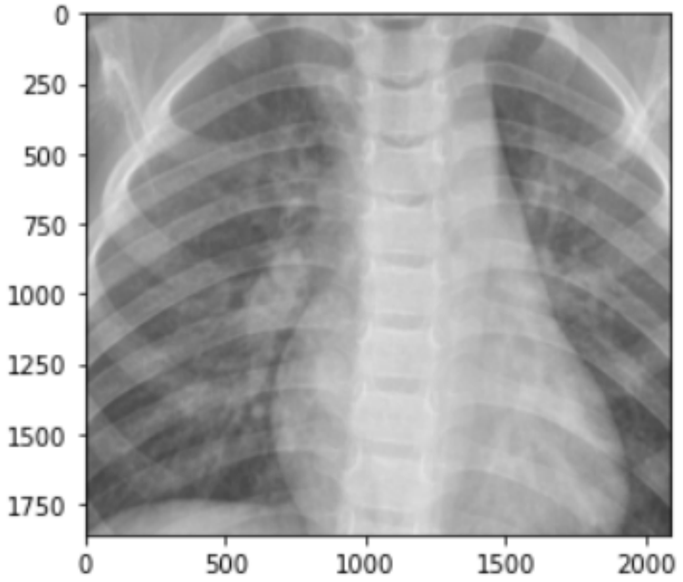
EXPLORATION DES DONNÉES

Traitement des données









Les différents models explorés

Présentation des différentes méthodes de ML utilisées

LES DIFFÉRENTS MODEL EXPLORÉS

Présentation

- Logistic Regression
- XGBoost
- CATBoost
- Neural network
- Optimisation bayésienne
- Transfert de connaissances

LES DIFFÉRENTS MODEL EXPLORÉS

Logistic Regression

- Choisi pour sa rapidité
- Bonne base de prédiction

LES DIFFÉRENTS MODEL EXPLORÉS

Gradient boosting

- Bon compromis entre les solutions basiques et celles plus gourmandes en ressources

- Facilement interprétable

LES DIFFÉRENTS MODEL EXPLORÉS

Neural Networks

- Plus complexe à implémenter
- Plus consommateurs en ressources machine
- Plus précis

LES DIFFÉRENTS MODEL EXPLORÉS

Optimisation Bayesienne

- Trouver automatiquement les meilleurs hyperparamètres

- Economise les ressources

LES DIFFÉRENTS MODEL EXPLORÉS

Fine tuning

- Utiliser un modèle pré-entrainé

- Gain de ressource et de précision

Analyse des résultats

Présentation des différentes résultats obtenus et comparaison des models de ML utilisés

Analyse des résultats

Metrics

- précision, recall, roc auc, score f1 ?

- le score ROC AUC "macro"

Tableau de conclusion

Nom du model	ROC AUC	Avantages	Inconniénients	Conclusion
Régression logistique	0.839	Fast	modèle linéaire	Pas pris en compte
XGBClassifier	0.907	Facilement interprétable et précis	Moyennement rapide	Pris en compte
CatBoostClassifier	0.906	Facilement interprétable et précis	Moyennement rapide	Pas pris en compte

Tableau de conclusion

Nom du model	ROC AUC	Avantages	Inconniénients	Conclusion
Réseau de neurones (SeNet)	0.904	Précis	Pas facilement interprétable/long à entrainer	Pas pris en compte
Réseau de neurones (transmet de connaissance)	0.936	Très précis et rapide à entrainer	Pas facilement interprétable	Pris en compte

Matrice des métriques et de confusion

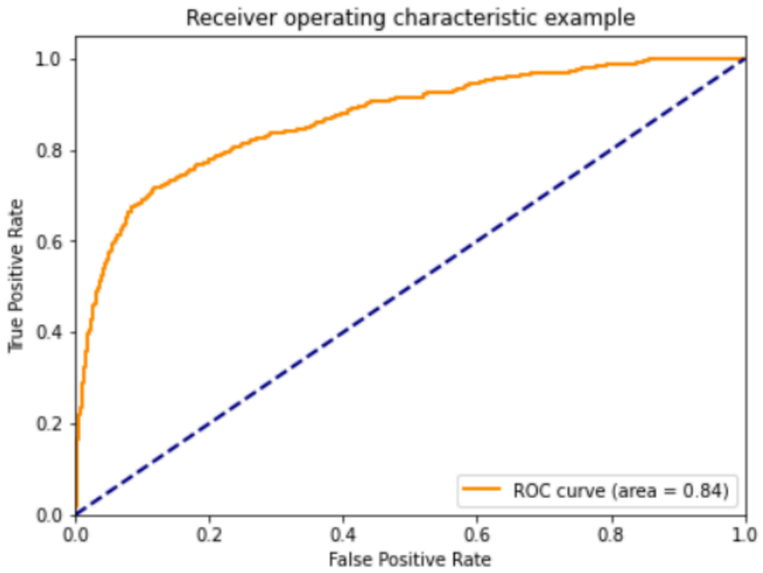
Analyse détaillée des résultats

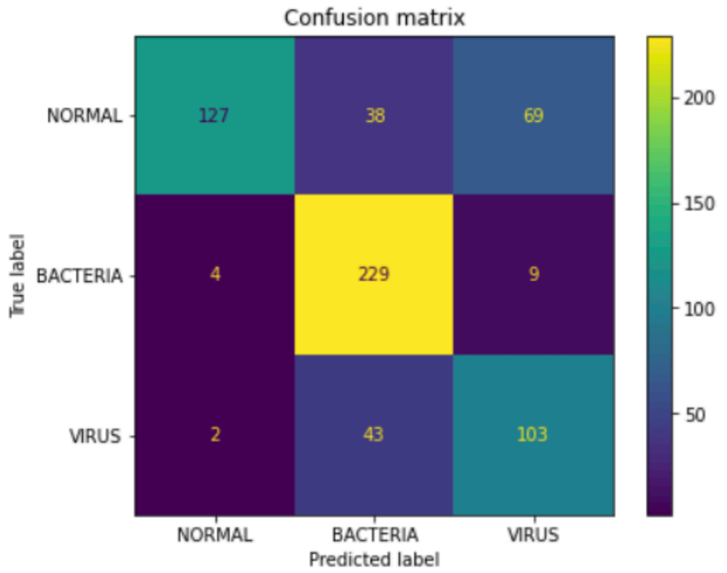
Matrice des métriques et de confusion

XGBClassifier

classification report:

	precision	recall	f1-score	support
NORMAL	0.95	0.54	0.69	234
BACTERIA	0.74	0.95	0.83	242
VIRUS	0.57	0.70	0.63	148
accuracy			0.74	624
macro avg	0.75	0.73	0.72	624
weighted avg	0.78	0.74	0.73	624





Matrice des métriques et de confusion

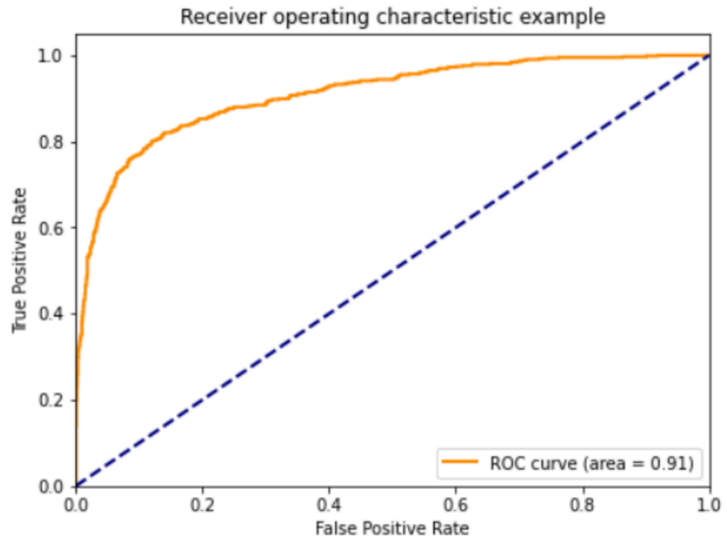
Réseau de neurone

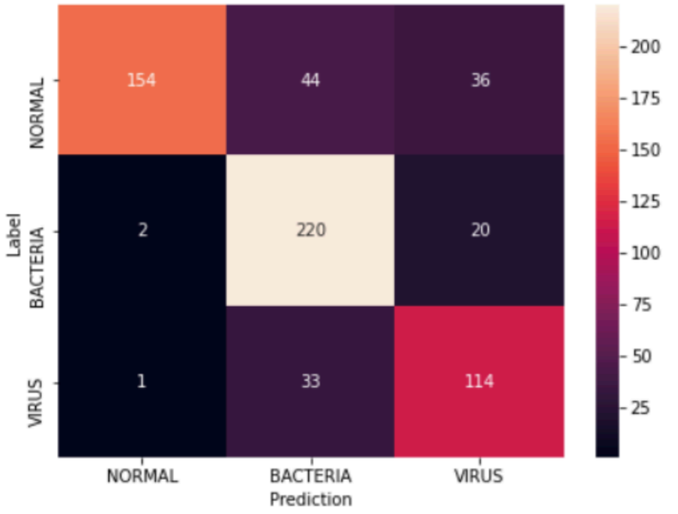
Voici nos résultats pour notre réseau de neurones:

```
classification report:
              precision    recall  f1-score   support

      NORMAL       0.98      0.66      0.79       234
    BACTERIA       0.74      0.91      0.82       242
       VIRUS       0.67      0.77      0.72       148

    accuracy                           0.78       624
   macro avg       0.80      0.78      0.77       624
weighted avg       0.81      0.78      0.78       624
```



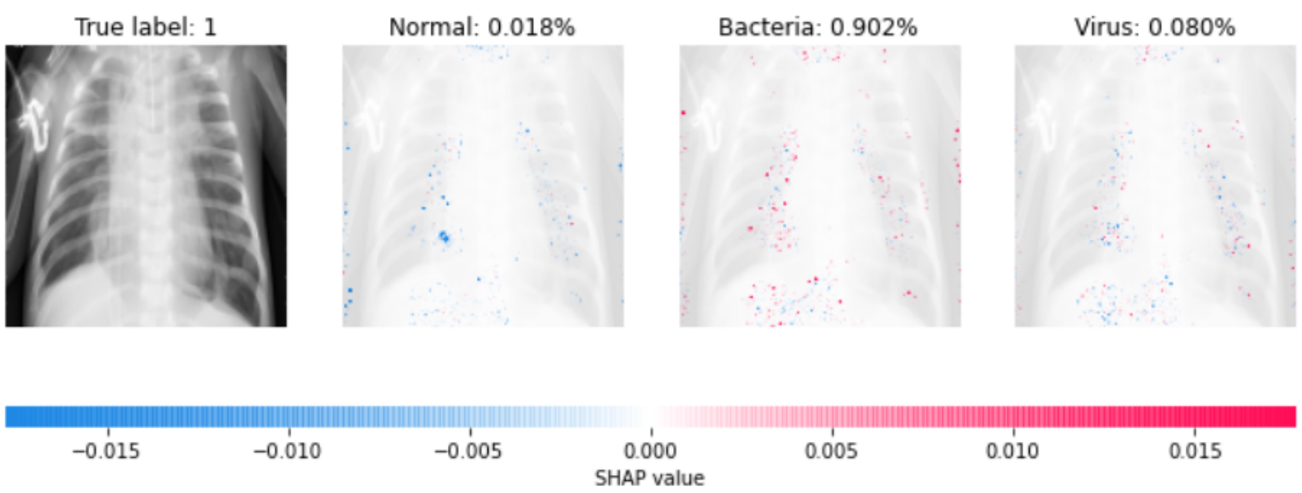


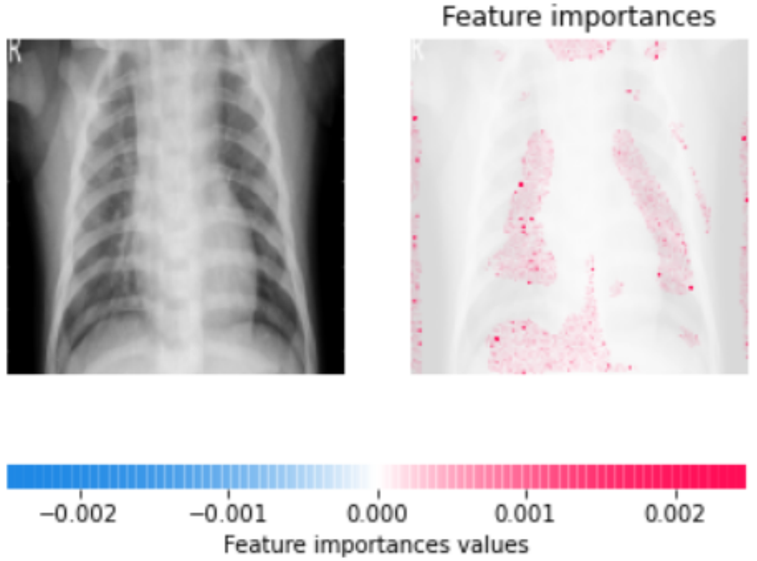

```
macro ROC AUC: 0.936
```

Interprétabilité des résultats

Interprétabilité des résultats

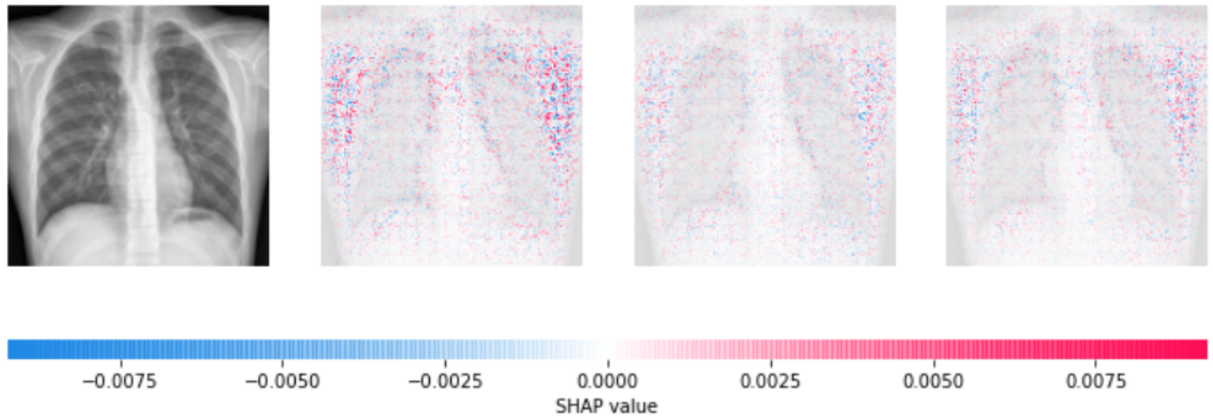
XGBClassifier

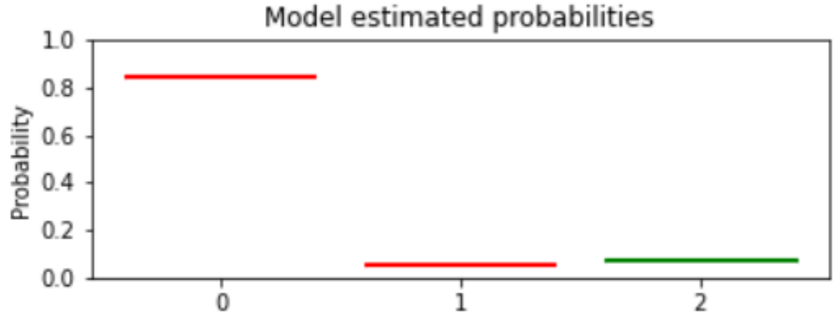




Interprétabilité des résultats

Réseau de neurone



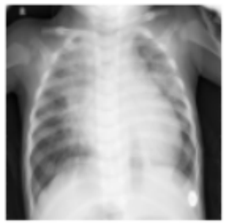


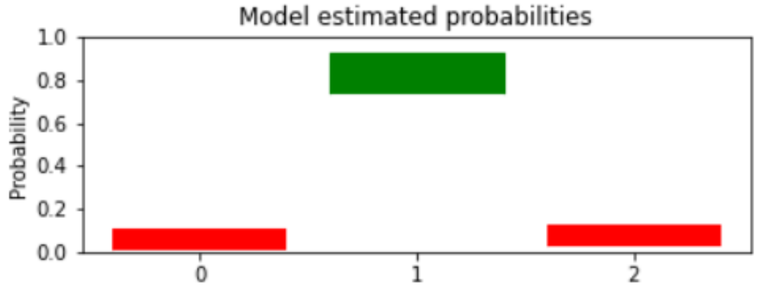
Models probabilistes

Models probabilistes

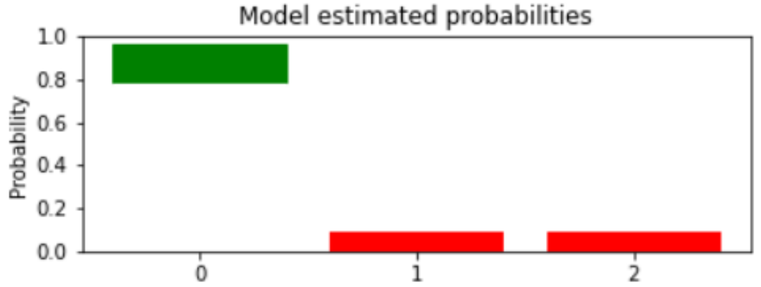
XGBClassifier + bagging

True label: 1

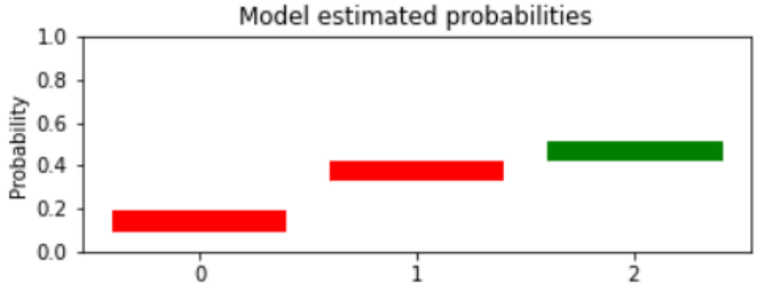




True label: 0

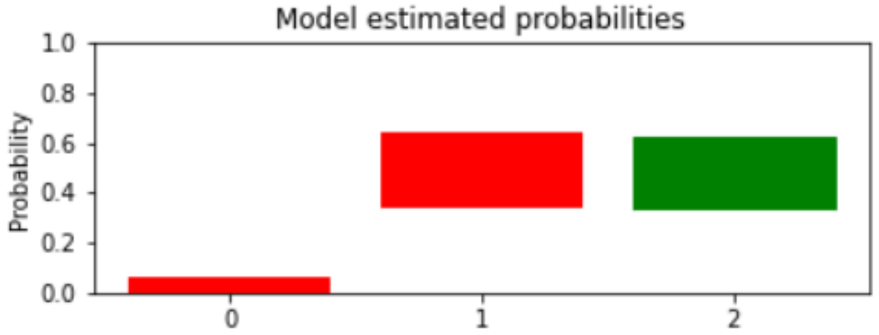


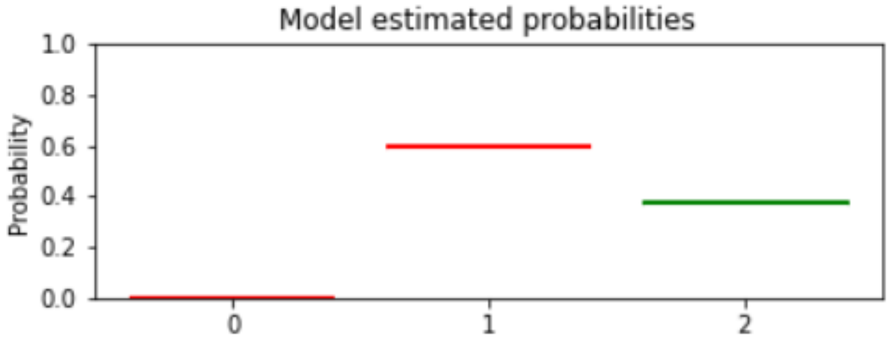
True label: 2



Models probabilistes

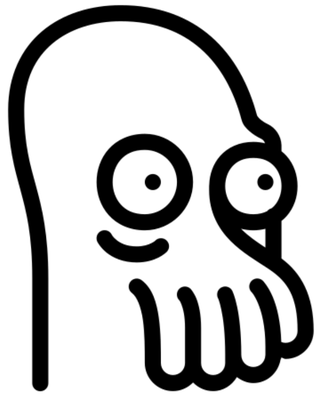
Réseau de neurone





Conclusion

- Le MobileNetV2 plus performant
- Models classiques plus facilement interprétables



Merci!

Avez vous des questions?