Chapitre 3

ANALYSE ET ELABORATION DU MODELE DE PREDICTION DE CANCER DU SEIN

Pour ce chapitre, il sera question de présenter les données, les résultats fournis par le modèle, ainsi qu'une analyse sur sa performance.

1.1 Introduction

L'étude des masses mammaires et leur classification dans le cadre du cancer du sein, ainsi que l'étude de l'état de l'art concernant les différents systèmes de diagnostic assisté par ordinateur.

L'étude de ces systèmes nécessite d'abord la comparaison de la performance de différents classifieurs à savoir la régression logistique (RL), le réseau de neurones en fonctions de base radiales (RBF), le réseau de neurones à convolution (CNN) et les séparateurs à vaste marge (SVM). C'est sur base de ce qui précède que nous avons choisi le CNN comme modèle de prédiction.

1.2 Présentation des données

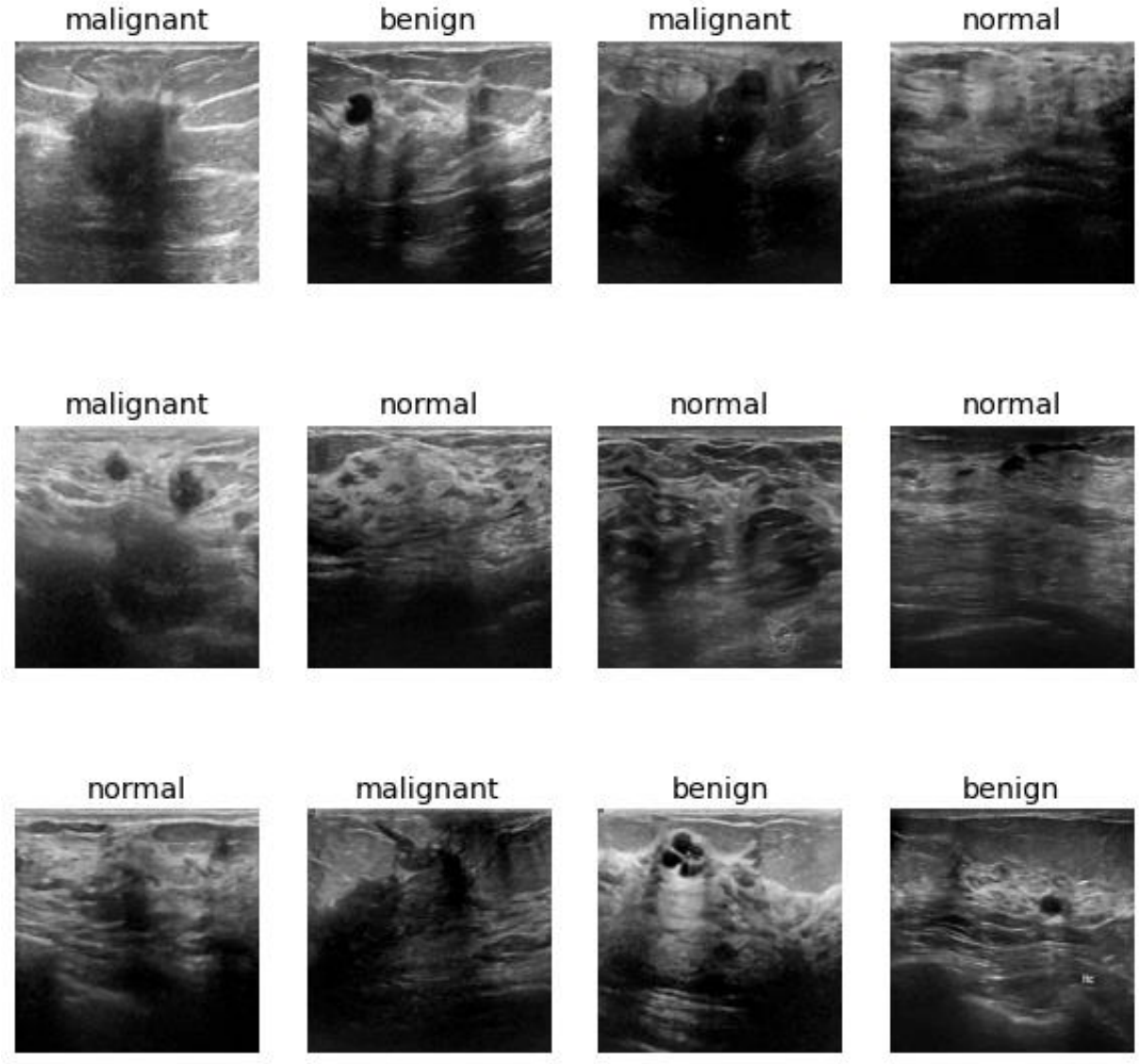
Rappelons que les données qui ont servies d'analyse et d'élaboration du modèle, sont des images.

Sur le plan médical et analytique, il existe certaines règles et consignes à respecter avant toute chose. Parmi eux nous pourrons en citer :

- La préservation du secret et de la vie privée des patients
- La base à tester doit être riche et doit comprendre tous les cas de figure possibles
- La base doit être connue par les chercheurs afin de faciliter la tâche de comparaison avec les travaux antérieurs.
- Etc.

Graphiquement les images se présentent comme :

Graphique 1 : Images échographiques de cancer du sein



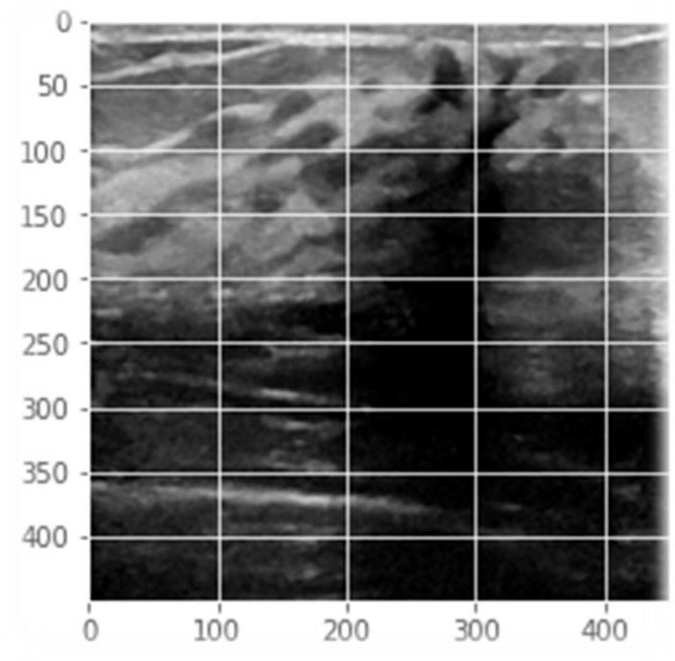
Ces images représentent les trois catégories des images qui ont servi à l'élaboration du modèle.
Source : A l'aide du langage de programmation Python 3.9.6

Dans le cas de cette analyse, les images seront considérées comme exogène et le prédicteur sera sous la forme catégoriel. Pour l'analyse, 1 représentera l'image bénigne, 2 représentera l'image maligne et 3 pour l'image normale.

Rappelons qu'une image est définie par les pixels[1] qui la renferme et le nombre de couleur[2] qui la compose. Chaque couleur qui compose une image est prise sous une échelle de 0 à 255 d'après normes actuelles de l'informatique.

La représentation graphique d'une image sous un système d'axe est :

Graphique 2 : Représentation d'une image sur un axe



Source : A l'aide du langage de programmation Python 3.9.6

[1] **Pixel** : Pour dire simple, le pixel correspond à l'intersection d'une ligne et d'une colonne dans un tableau.
[2] **Nombre de couleur qui la compose** : Une image dite **en couleur** est celle qui est principalement composée de la combinaison de trois couleurs. Les plus connues sont le rouge, le vert et le bleu (RGB).

1.3 Résultats du modèle

Rappelons que le modèle utilisé pour cette étude est le réseau de neurones à convolution. Les résultats se présentent comme :

Figure 1 : Extrait des sorties du modèle

Model: "sequential_2"			
_____			
Layer (type)	Output Shapxe	Param #	
==========			
sequential (Sequential)	(16, 450, 450, 3)	0	
_____			
sequential_1 (Sequential)	(16, 450, 450, 3)	0	
_____			
conv2d (Conv2D)	(16, 448, 448, 32)	896	
_____			
max_pooling2d (MaxPooling2D)	(16, 224, 224, 32)	0	
_____			
conv2d_1 (Conv2D)	(16, 222, 222, 64)	18496	
_____			
max_pooling2d_1 (MaxPooling2	(16, 111, 111, 64)	0	

==========

Source : A l'aide du langage de programmation Python 3.9.6

Total params : 269,763

Trainable params : 269,763

Non-trainable params : 0

Le modèle comporte 269763 paramètres, ce qui pose une petite difficulté dans l'écriture. Mais il est possible de sauvegarder le modèle sous la forme d'un fichier (avec une extension .h5, .pkl, .tflite, …).

Les images ont été regroupées en 16 groupes, chaque image a un format de 450*450 pixels et à trois couleurs.

N.B : beaucoup de détails sur le modèle seront présentés dans les points qui vont suivre.

1.4 Evaluation du modèle

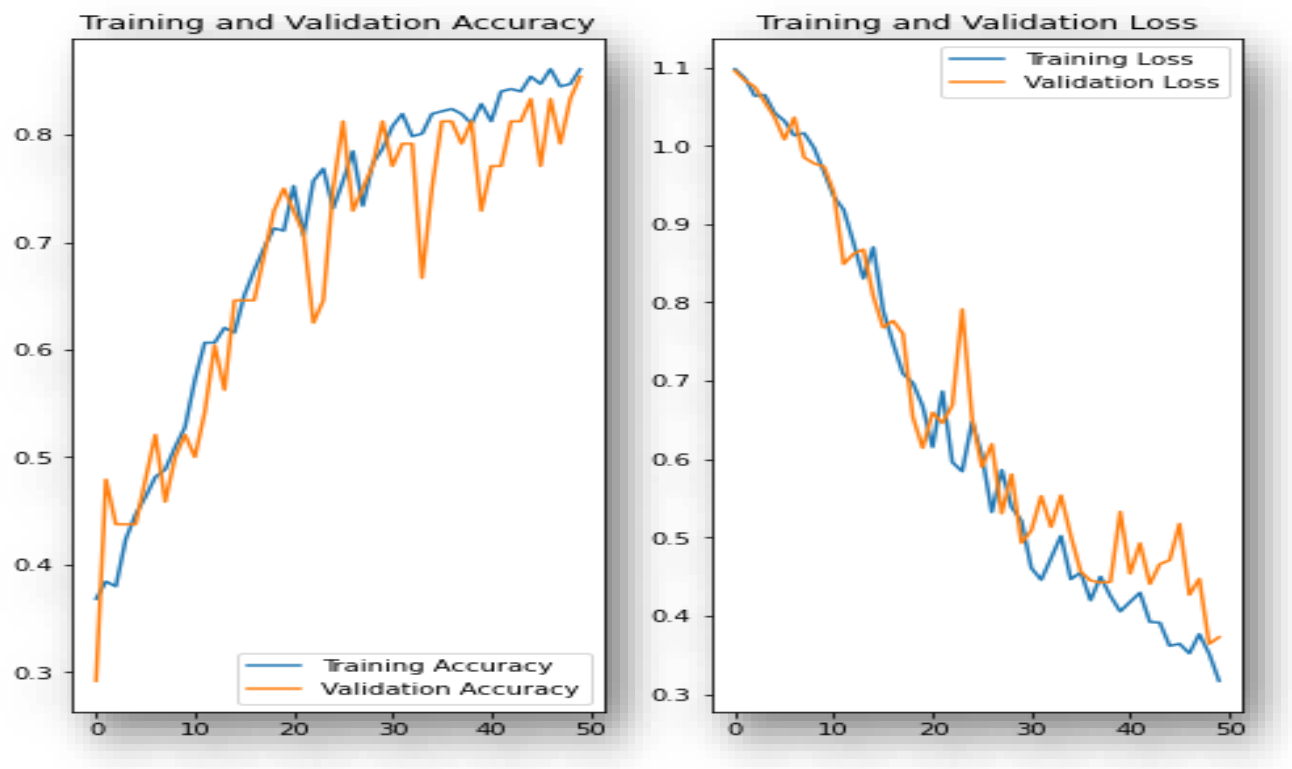
Tableau 1 : Evaluation du modèle sur le train, test et validation

Train	Validation	Test
0.8611	0.8542	0.8594
Source : A l'aide du langage de programmation Python 3.9.6		

Globalement le modèle est de très bonne qualité. Le modèle est capable de prédire à 86% sur les données qui ont servi de création du modèle. Sur les données qui ont servi de validation, le modèle est capable de prédire à 85% et sur les données de test, le modèle est capable de prédire à 86%. Tout ce qui précède signifie que si nous présentons une nouvelle image au modèle, notre modèle aura une probabilité de 0.86 de réussir sa prédiction. Ce qui conduit à dire que la confiance sur le modèle ne sera pas décevante.

Graphiquement, cette performance est présentée comme :

Graphique 3 : Apprentissage du modèle sur les données train et de validation



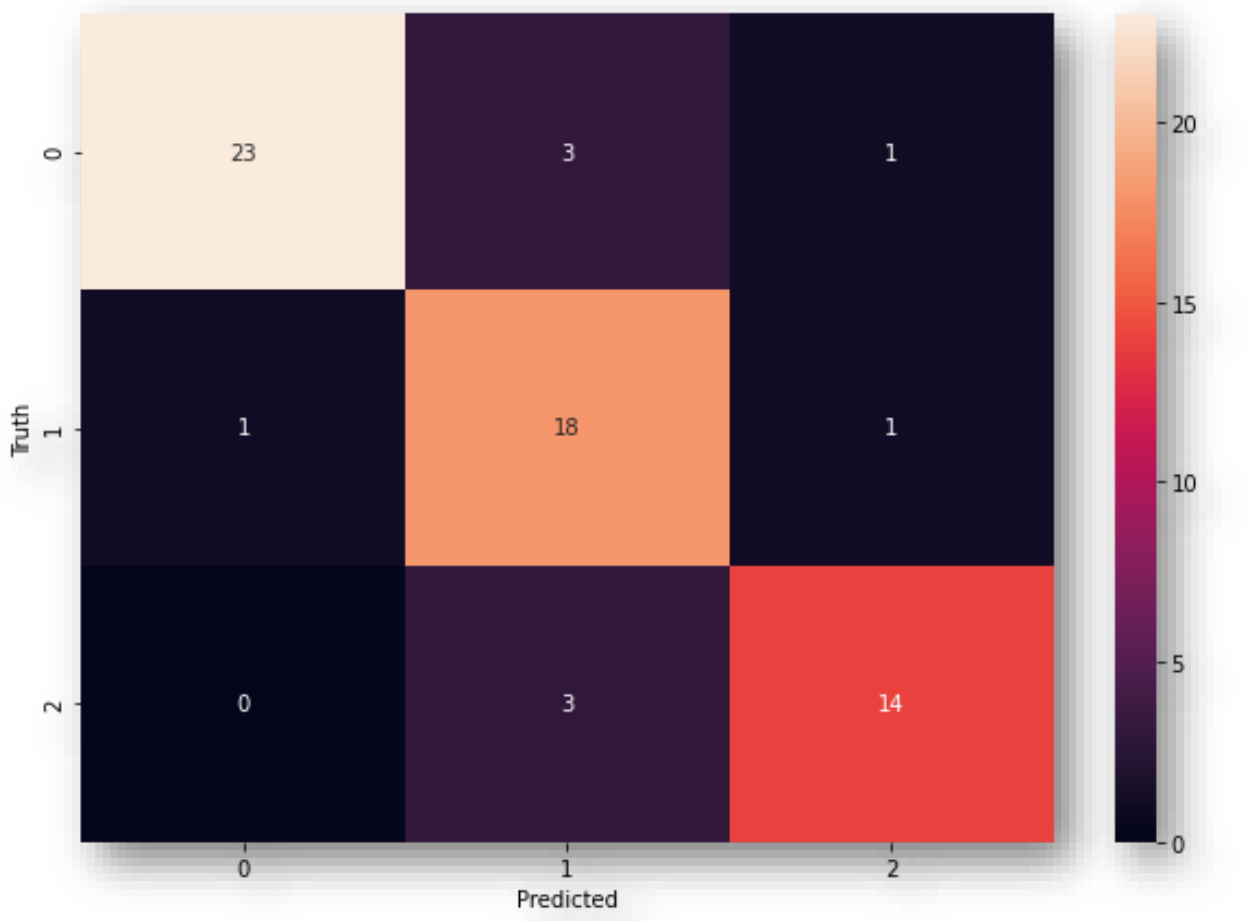
Source : A l'aide du langage de programmation Python 3.9.6

1.5 Performance du modèle

Pour mesurer la performance du modèle, il serait question de se servir de la matrice de la matrice de confusion (avec trois classes).

Cette matrice de confusion se présente comme :

Graphique 4 : Matrice de confusion



Source : A l'aide du langage de programmation Python 3.9.6

Tableau 2 : Résumé des résultats de la matrice de confusion

	precision	recall	f1-score	support
benign	0.96	0.85	0.90	27
malignant	0.75	0.90	0.82	20
normal	0.88	0.82	0.85	17
accuracy			0.86	64
macro avg	0.86	0.86	0.86	64
weighted avg	0.87	0.86	0.86	64

Source : A l'aide du langage de programmation Python 3.9.6

1.5.1 Sensibilité

Rappelons que la sensibilité (ou rappel) qui est, en effet, le taux des vrais positifs (TVP), représente la capacité du modèle à retrouver les positifs, càd la capacité d'un examen diagnostique à fournir un résultat positif en présence de la maladie.

Elle se calcul de la manière suivante :

$$TVP = Sensibilité = \frac{VP}{VP + FN}$$

Avec :

VP (Vrai Positif) : est le nombre de lésions malignes qui sont classées malignes.
FP (Faux Positif) : est le nombre de lésions bénignes (ou normales) qui sont classées malignes.
VN (Vrai négatif) : est le nombre de lésions bénignes (ou normales) qui sont classées bénignes (ou normales).
FN (Faux Négatif) : est le nombre de lésions malignes qui sont classées bénignes ou normales.

Rappelons que la lésion maligne constitue le cas positif pour cette étude, car celle qui constitue le cancer de sein.

$$S_e = \frac{18}{1+18+1} = 0.90$$

Pour ce modèle, la capacité d'un examen diagnostique à fournir un résultat de lésion maligne en présence de la maladie est estimée à 90%.

1.5.2 Taux d'erreur

Le taux d'erreur est égal au nombre de mauvais classement rapporté à l'effectif total.

$$\varepsilon = \frac{3+1+1+1+0+3}{64} = 0.14$$

La probabilité de mauvais classement est estimée à 0.14 ou à 14%. Cela signifie que si nous sommes en présence d'une lésion maligne, la probabilité pour que nous ayons une lésion bénigne (respectivement normale) est estimée à 14%.

1.5.3 Taux de succès

Le taux de succès est égal au nombre de bon classement rapporté à l'effectif total, ou même le complémentaire à l'unité du taux d'erreur.

$$\theta = 1 - \varepsilon = \frac{23+18+14}{64} = 1 - 0.14 = 0.86$$

La probabilité de bon classement est estimée à 0.86 ou à 86%. Cela signifie que si nous sommes en présence d'une lésion maligne, la probabilité pour que nous ayons une lésion maligne est estimée à 86%.

1.5.4 Spécificité

La spécificité, à l'inverse de la sensibilité, indique la proportion des négatifs détectés.

Pour cette étude, la lésion bénigne et le cas normal constituent le cas négatif. En rapport avec ce qui vient d'être dit, le calcul de la spécificité se fera de deux manières (pour le cas normal et pour la lésion bénigne).

$$Sp_{Bénigne} = \frac{23}{23+3+1} = 0.85$$

$$Sp_{Normal} = \frac{14}{0+3+14} = 0.82$$

Ce modèle a une capacité de 85% à pouvoir diagnostiquer la lésion bénigne et celle 82% à pouvoir diagnostiquer le cas normal.

1.5.5 Précision

La précision indique la proportion de vrais positifs parmi les individus qui ont été classés positifs. Càd la proportion des vrais lésions malignes parmi les lésions malignes classées.

$$P = \frac{18}{3+18+3} = 0.75$$

La proportion des vrais lésions malignes parmi les lésions malignes classées est de 75%.

1.5.6 F-Mesure (ou f1-score)

Ce n'est rien d'autre que la moyenne harmonique entre la précision et la sensibilité. Elle synthétise la précision et la sensibilité.

A ce niveau, on accorde la même importance au rappel et à la précision, la F-Mesure devient :

Pour la lésion maligne, nous avons $f1-score = \frac{2}{\frac{1}{S_e}+\frac{1}{P}} = \frac{2*P*S_e}{P+S_e} = \frac{2*0.75*0.90}{0.75+0.90} = 0.82$

1.5.7 Rapport de vraisemblance

Le rapport de vraisemblance décrit le surcroît de chances des positifs (par rapport aux négatifs) d'être classés positifs. Sa définition est la suivante :

$$L = \frac{S_e}{1-Sp}$$

$$L_{Bénigne} = \frac{0.9}{1-0.85} = 6$$

$$L_{Normal} = \frac{0.9}{1-0.82} = 5$$

Pour une personne diagnostiquée, la chance d'être diagnostiquer maligne est 6 fois de plus grandes que celle d'être bénigne et 5 fois plus grande que d'être normal.

1.5.8 Pseudo-R^2 basé sur le taux d'erreur

Avant de pouvoir calculer d'abord le pseudo-R^2, il serait nécessaire de déterminer la matrice de confusion du classifieur par défaut.

La règle de décision du classifieur par défaut est donc très simple : on affecte, pour tout individu à classer, la modalité majoritaire dans l'échantillon d'apprentissage.

La matrice de confusion du classifieur par défaut se présente comme :

Tableau 3 : Matrice de confusion du classifieur par défaut

	0	1	2
0	0	0	23
1	0	0	18
2	0	0	14

Source : A l'aide du langage de programmation Python 3.9.6

Et le taux d'erreur associé est :

$$\varepsilon(def) = \frac{23+18}{55} = 0.74545$$

Le taux d'erreur par défaut est :

$$\varepsilon(M) = 0.14$$

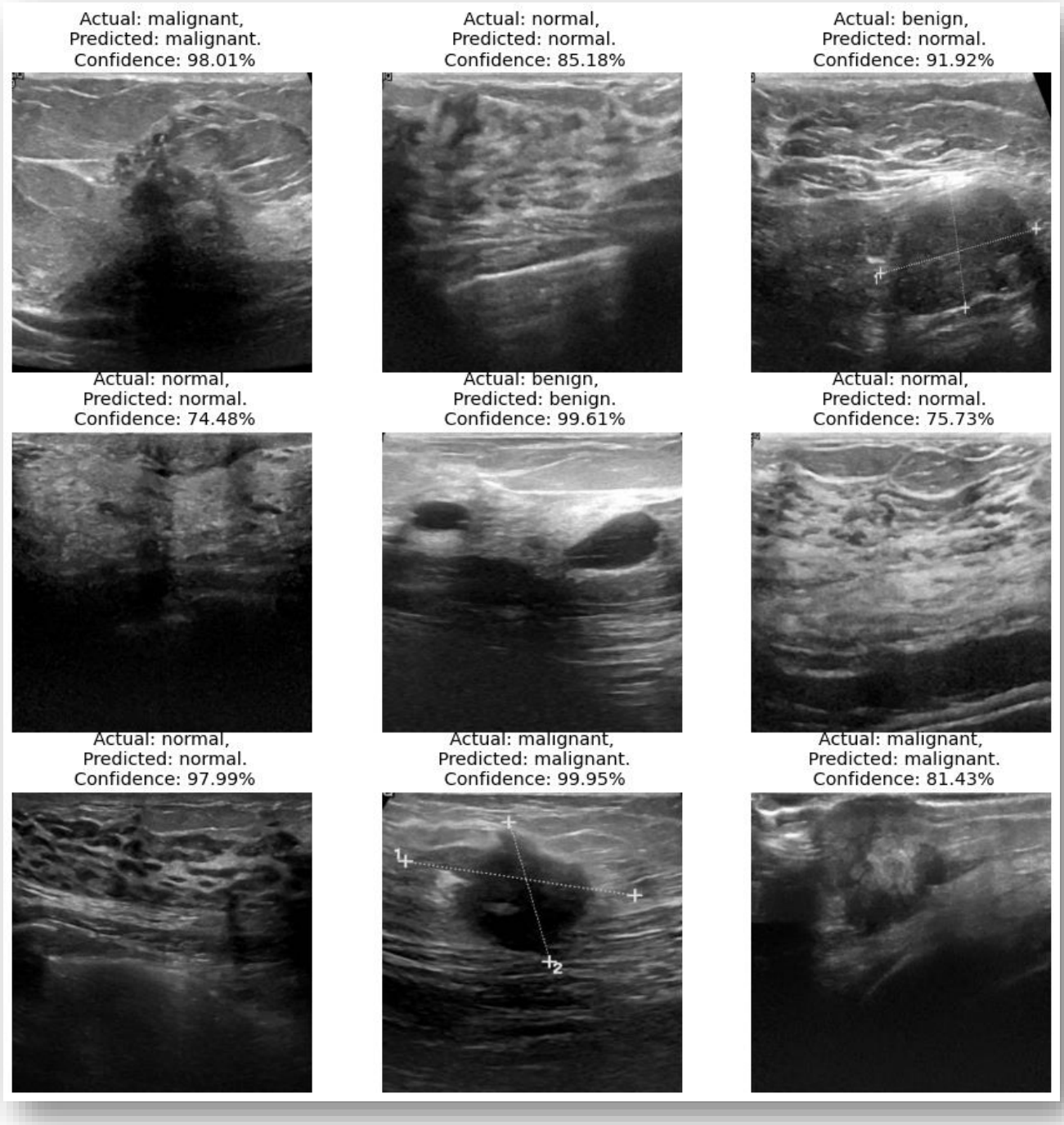
De ce fait, le pseudo-R^2 vaut :

$$R_{\varepsilon}^2 = 1 - \frac{\varepsilon(M)}{\varepsilon(def)} = 1 - \frac{0.14}{0.74545} = 0.81$$

Le modèle servant à cette étude est mieux que le classifieur par défaut, d'où la confiance au modèle est accrue.

1.6 Prédiction

Graphique 5 : Prédiction avec probabilité de succès



Source : A l'aide du langage de programmation Python 3.9.6

1.7 Conclusion

Nous voici à terme de ce chapitre qui traite de l'analyse allant de la présentation des données, jusqu'à la prédiction, et de l'élaboration du modèle de prédiction de cancer du sein. Sur ce, nous sommes passé en revue des étapes suivantes :

- Présentation des données
- Présentation du résultat
- Évaluation du modèle
- Mesure de la performance du modèle
- Prédiction

Partant des étapes qui précèdent, le modèle s'est révélé confiant à 85% au moins sur toute catégorie des données (train, test et validation). C'est ce qui est permis de pouvoir aller aux prédictions (ce qui serait appelé en médecine comme diagnostic).