

# Rapport sur le Projet: Classification d'Images pour la Détection du Cancer du Sein

<https://www.linkedin.com/in/alioune-gaye-1a5161172/>

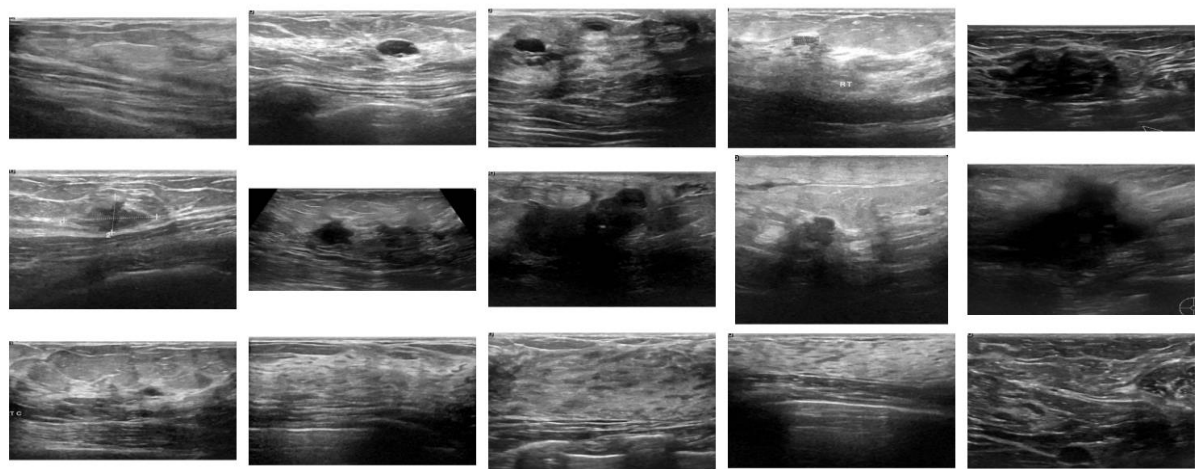
## 1. Introduction

Le cancer du sein est une des principales causes de mortalité chez les femmes. Ce projet exploite des images échographiques mammaires pour classer les cas en trois catégories principales : bénin, malin, et normal. Deux modèles d'apprentissage profond, VGG16 et Xception, ont été utilisés pour évaluer la performance et l'efficacité des approches dans ce contexte.

## 2. Aperçu du jeu de données

	image_path	label
0	Dataset_BUSI_with_GT\benign\benign (1).png	benign
1	Dataset_BUSI_with_GT\benign\benign (10).png	benign
2	Dataset_BUSI_with_GT\benign\benign (100).png	benign
3	Dataset_BUSI_with_GT\benign\benign (101).png	benign
4	Dataset_BUSI_with_GT\benign\benign (102).png	benign

5 Sample Images from Each Category



## 2.1 Description des données

Le jeu de données contient des images échographiques associées à trois catégories :

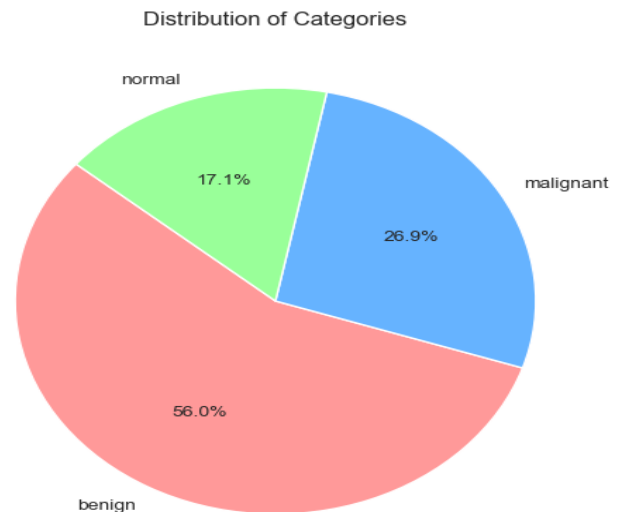
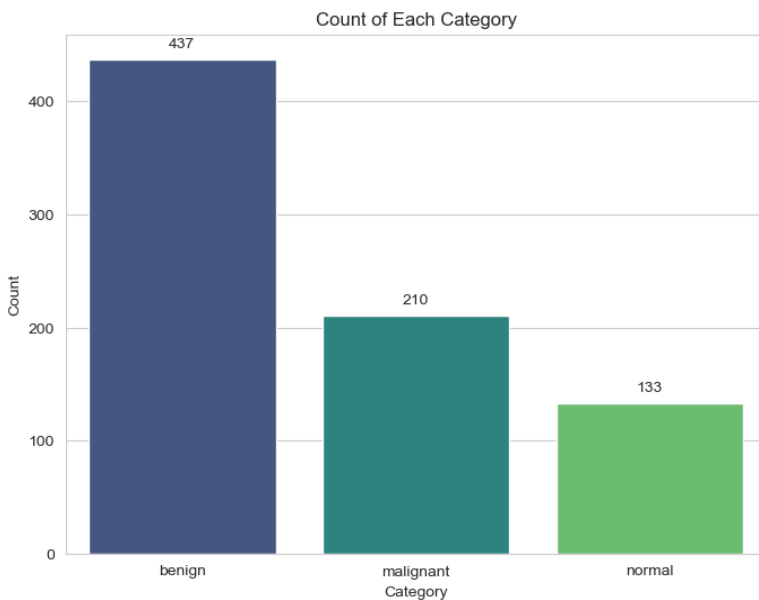
- **Bénin** : Masse détectée mais non cancéreuse.
- **Malin** : Masse détectée et identifiée comme cancéreuse.
- **Normal** : Aucun signe d'anomalie détecté.

## 2.2 Distribution initiale des catégories

Avant toute modification, les données étaient fortement déséquilibrées :

- **Bénin** : 437 images (56 %). La majorité des images sont associées à cette catégorie, ce qui reflète un déséquilibre important dans le jeu de données.
- **Malin** : 210 images (27 %). Cette catégorie est sous-représentée, bien qu'elle soit critique pour une détection précise.
- **Normal** : 133 images (17 %). La classe la moins représentée.

Cet important déséquilibre pourrait biaiser les performances des modèles, rendant difficile la détection des cas malins et normaux. Une méthode de sur-échantillonnage sera appliquée pour équilibrer les données.



## 2.2 Préparation des données

- **Redimensionnement des images** : Toutes les images ont été redimensionnées à 224x224 pixels.
- **Encodage des étiquettes** : Les catégories textuelles ont été converties en valeurs numériques (bénin = 0, malin = 1, normal = 2).

Class distribution after oversampling:

category\_encoded

0 437

1 437

2 437

Name: count, dtype: int64

	image_path	category_encoded
0	Dataset_BUSI_with_GT\benign\benign (1).png	0
1	Dataset_BUSI_with_GT\benign\benign (10).png	0
2	Dataset_BUSI_with_GT\benign\benign (100).png	0
3	Dataset_BUSI_with_GT\benign\benign (101).png	0
4	Dataset_BUSI_with_GT\benign\benign (102).png	0
...	...	...
1306	Dataset_BUSI_with_GT\normal\normal (76).png	2
1307	Dataset_BUSI_with_GT\normal\normal (60).png	2
1308	Dataset_BUSI_with_GT\normal\normal (6).png	2
1309	Dataset_BUSI_with_GT\normal\normal (115).png	2
1310	Dataset_BUSI_with_GT\normal\normal (133).png	2

1311 rows x 2 columns

- **Gestion du déséquilibre des classes** : Une méthode de sur-échantillonnage a été utilisée pour équilibrer les données.

### 2.3 Division des ensembles

Les données ont été divisées en :

- Entraînement : 80 % (1 048 images).
- Validation : 10 % (131 images).
- Test : 10 % (132 images).

## 3. Méthodologie

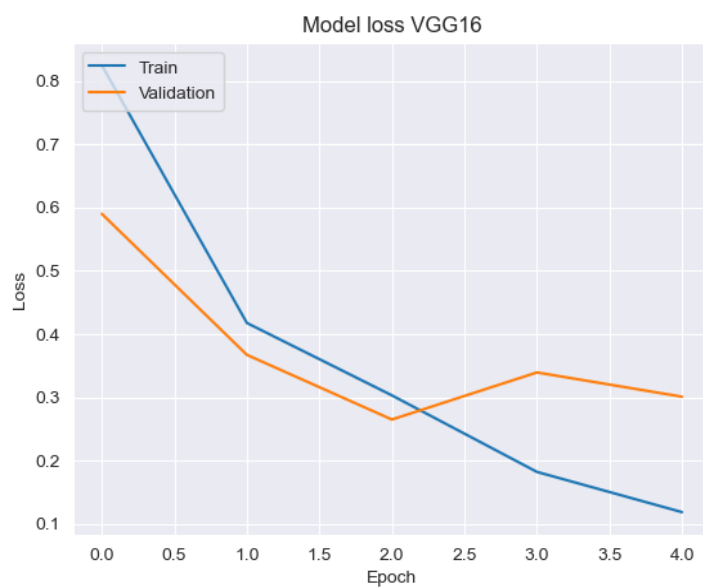
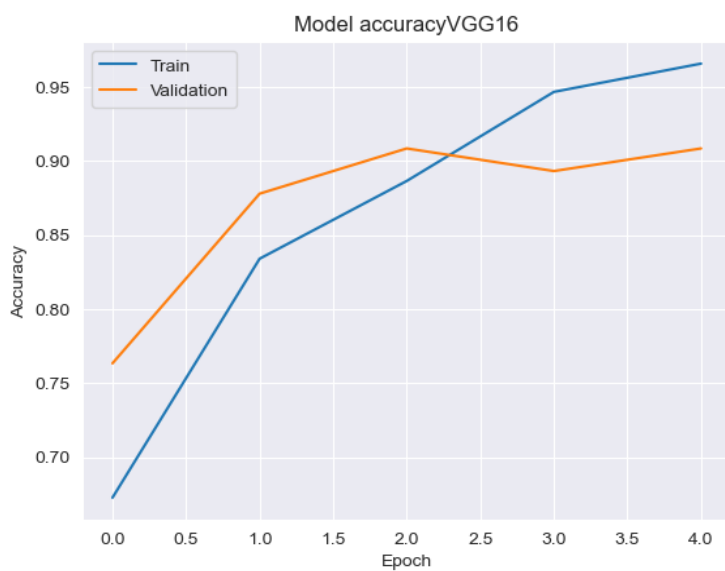
### 3.1 Modèles utilisés

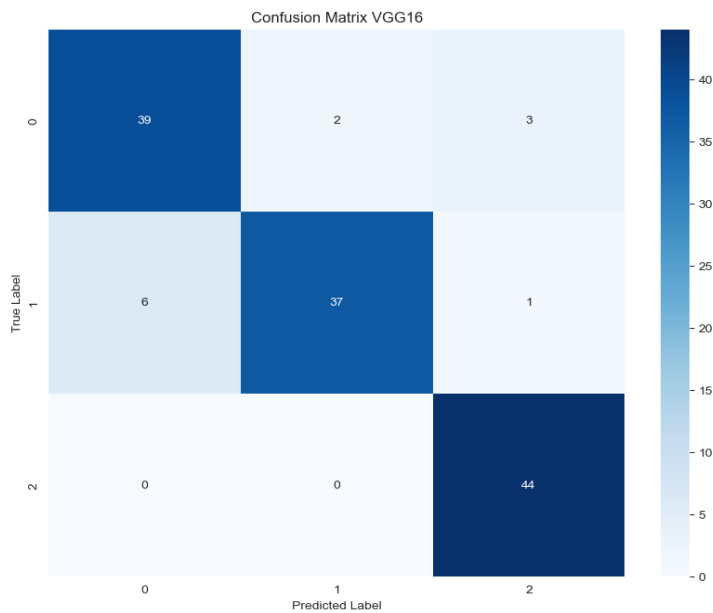
- VGG16 modifié : Ajout d'une attention multi-tête et régularisation avec Dropout et bruit gaussien.
- Xception modifié : Architecture avancée intégrant une attention multi-tête et des couches de régularisation.

### 3.2 Entraînement

- Optimiseur : Adam avec un taux d'apprentissage de 0.0001.
- Perte : Sparse Categorical Crossentropy.
- Métriques suivies : Précision et perte pour l'entraînement et la validation.
- Nombre d'époques : 5 (avec arrêt anticipé).

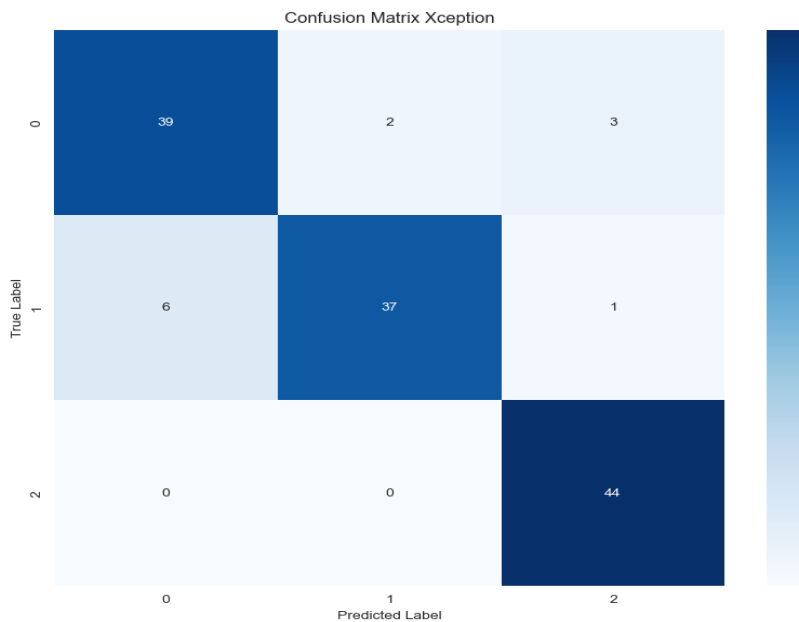
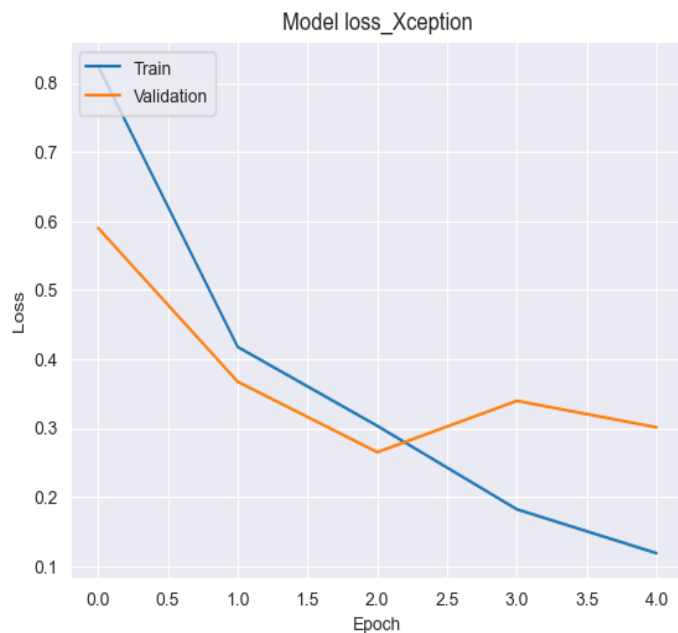
## 4. Résultats





report\_VGG16

	precision	recall	f1-score	support
0	0.81	0.68	0.74	44
1	1.00	0.14	0.24	44
2	0.49	1.00	0.66	44
accuracy			0.61	132
macro avg	0.77	0.61	0.55	132
weighted avg	0.77	0.61	0.55	132



report\_Xception

	precision	recall	f1-score	support
0	0.87	0.89	0.88	44
1	0.95	0.84	0.89	44
2	0.92	1.00	0.96	44
accuracy			0.91	132
macro avg		0.91	0.91	132
weighted avg		0.91	0.91	132

## 5. Comparaison entre VGG16 et Xception

Critères	Modèle VGG16	Modèle Xception
Précision moyenne	61 %	91 %
Précision pour classe 0	81 %	87 %
Précision pour classe 1	100 %	95 %
Précision pour classe 2	49 %	92 %
Rappel global	61 %	91 %
Généralisation	Moyenne (légère tendance à overfit)	Excellente (stabilité entre validation et entraînement)
Avantages principaux	Simplicité et rapidité relative	Performance globale exceptionnelle
Limites	Faibles performances globales malgré une précision élevée pour classe 1	Temps d'entraînement plus élevé

## 6. Conclusion et Perspectives

**Xception** surpasse **VGG16** en termes de généralisation et de précision globale, en particulier pour les classes difficiles à distinguer. Le choix final dépendra des ressources disponibles, mais pour des diagnostics cliniques précis, Xception est recommandé.

### Perspectives

Collaboration avec des instituts de recherche et des centres cliniques, pourrait accélérer la validation clinique de mes outils, tout en intégrant des solutions dans des essais cliniques novateurs pour des traitements ciblés, comme l'immunothérapie combinée. Ces améliorations offrent une perspective multidimensionnelle qui renforcerait l'efficacité de ce projet pour une prise en charge complète et individualisée des cancers du sein, du diagnostic précoce au suivi thérapeutique.