amU
Aix Marseille Université

NNT : 2024AIXM0591

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 10 décembre 2024 par

Jules COLLENNE

Aide au diagnostic du mélanome : Approche par apprentissage de similarité

Discipline
Informatique

École doctorale
184 MATHEMATIQUES ET INFORMATIQUE

Laboratoire/Partenaires de recherche
Laboratoire d'Informatique et des Systèmes (LIS)

Composition du jury

Laurent HEUTTE PR, Université Rouen Normandie	Président du jury, Rapporteur
Christine FERNANDEZ-MALOIGNE PR, Université de Poitiers	Rapporteure
Emanuel ALDEA MCF, Université Paris-Saclay	Examinateur
Caroline GAUDY PU-PH, Service de Dermatologie de l'APHM	Examinatrice
Monique THONNAT DR, INRIA Sophia Antipolis	Examinatrice
Rabah IGUERNAISSI MCF, LIS, Aix-Marseille Université	Co-encadrant de thèse
Séverine DUBUISSON PR, LIS, Aix-Marseille Université	Co-directrice de thèse
Djamal MERAD PR, LIS, Aix-Marseille Université	Directeur de thèse

LIS
LABORATOIRE D'INFORMATIQUE & DES SYSTÈMES

Affidavit

Je soussigné, Jules Collenne, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Djamal Merad et Séverine Dubuisson, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte AMU relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 16 septembre 2024



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

1. *(En cours de soumission)* **J. Collenne**, R. Iguernaissi, S. Dubuisson, D. Merad. “SJ-MAE : Learning Visual Representations with Siamese Masked Autoencoders and Jigsaw Puzzles”
2. **J. Collenne**, R. Iguernaissi, S. Dubuisson, D. Merad. “ReSet : A Residual Set-Transformer approach to tackle the ugly-duckling sign in melanoma detection”. International Conference of Image Processing **(ICIP 2024)**
3. **J. Collenne**, R. Iguernaissi, S. Dubuisson, and D. Merad, "Enhancing Anomaly Detection in Melanoma Diagnosis Through Self-Supervised Training and Lesion Comparison," Machine Learning in Medical Imaging **(MLMI 2023)**
4. **J. Collenne**, Jilliana Monnier, Rabah Iguernaissi, Motasem Nawaf, Marie-Aleth Richard, Jean-Jacques Grob, Caroline Gaudy-Marqueste, Séverine Dubuisson, Djamal Merad, "Fusion between an Algorithm Based on the Characterization of Melanocytic Lesions’ Asymmetry with an Ensemble of Convolutional Neural Networks for Melanoma Detection", **Journal of Investigative Dermatology, 2024**, ISSN 0022-202X,https ://doi.org/10.1016/j.jid.2023.09.289.
5. A. C. Foahom Gouabou, **J. Collenne**, J. Monnier, R. Iguernaissi, J.-L. Damoiseaux, A. Moudafi, D. Merad, "Computer Aided Diagnosis of Melanoma Using Deep Neural Networks and Game Theory : Application on Dermoscopic Images of Skin Lesions." **International Journal of Molecular Sciences, 2022**, 23, 13838. https ://doi.org/10.3390/ijms23221383
6. J. Monnier, A. C. Gouabou Foahom, M. Serdi, **J. Collenne**, R. Iguernaissi, M.-A. Richard, C. Gaudy-Marqueste, J.-L. Damoiseaux, J.-J. Grob, D. Merad, "Automated melanoma detection. An algorithm inspired from human intelligence characterizing disordered pattern of melanocytic lesions improving a convolutional neural network.", **Journal of the American Academy of Dermatology, 2024**, ISSN 0190-9622, https ://doi.org/10.1016/j.jaad.2024.02.063.

Participation aux conférences et écoles d’été au cours de la période de thèse :

1. 2024 IEEE International Conference on Image Processing (ICIP 2024)

2. 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2023)
3. 2023 International Conference on Computer Vision (ICCV 2023)

Résumé

Le mélanome est un cancer de la peau dont la métastase rapide nécessite une prise en charge immédiate. Bien qu'il ne soit pas le cancer de la peau le plus fréquent, il demeure le plus mortel, et son incidence dans les pays développés ne cesse d'augmenter. Pour améliorer la prise en charge de cette maladie, la première étape consiste à accélérer le diagnostic, notamment en optimisant les méthodes d'imagerie et de diagnostic assisté par ordinateur. Les récentes avancées en intelligence artificielle et en apprentissage automatique ont déjà permis de créer de nombreux algorithmes capables de détecter diverses maladies, dont le mélanome. Ces approches pourraient, à terme, s'implanter dans le milieu médical afin d'assister les dermatologues dans l'analyse des lésions cutanées. Cette assistance pourrait rendre le diagnostic du mélanome plus efficace et précis, réduisant ainsi sa mortalité.

Plusieurs études montrent l'efficacité de ces méthodes. Toutefois, la plupart des travaux actuels n'utilisent pas efficacement toutes les connaissances dermatologiques, ce qui peut nuire à la qualité des prédictions des modèles. Cette thèse s'inscrit dans le besoin constant de lier les connaissances des dermatologues aux modèles d'apprentissage automatique. Pour ce faire, nous avons étudié en profondeur l'implémentation d'un concept dermatologique spécifique, celui du "Vilain Petit Canard". Ce concept met en avant la comparaison des lésions chez un même patient afin d'améliorer le diagnostic. L'idée générale de cette approche repose sur le fait que les lésions atypiques chez un même patient ont une probabilité plus élevée d'être cancéreuses.

Dans cette thèse, nous présentons différentes approches basées sur ce concept, établissant un lien entre les innovations en intelligence artificielle et les avancées médicales.

La première partie présente le contexte médical et scientifique du diagnostic du mélanome par vision par ordinateur.

La seconde partie présente deux approches que nous avons développées, qui s'attaquent à la tâche de comparaison intra-patient des lésions cutanées. L'une est une approche entièrement non supervisée, utilisant l'apprentissage de représentations visuelles et des algorithmes de détection d'anomalies sans annotations. L'autre est une approche supervisée, basée sur les réseaux de neurones et l'architecture des Transformers.

La troisième partie se concentre sur l'encodage non supervisé des images à l'aide de nouvelles méthodes d'apprentissage de représentations visuelles. Nous présentons les résultats de cette nouvelle méthode sur la classification du mélanome.

Enfin, la dernière partie traite de l'interprétabilité des modèles, notamment avec l'utilisation de modèles génératifs. Cette approche pourrait, à terme, améliorer l'interprétabilité des modèles d'intelligence artificielle en contexte clinique.

Mots clés : Intelligence artificielle, apprentissage profond, représentations visuelles, diagnostic du mélanome, transformers.

Abstract

Melanoma is a type of skin cancer that, due to its rapid metastasis, requires immediate medical attention. Although it is not the most common form of skin cancer, it remains the deadliest, and its incidence continues to rise in developed countries. To improve the management of this disease, the first step is to accelerate diagnosis, particularly by optimizing imaging methods and computer-assisted diagnostic tools. Recent advances in artificial intelligence and machine learning have already led to the creation of numerous algorithms capable of detecting various diseases, including melanoma. These approaches could eventually be integrated into the medical field to assist dermatologists in analyzing skin lesions. This assistance could make melanoma diagnosis more efficient and accurate, thus reducing its mortality rate.

Several studies have demonstrated the effectiveness of these methods. However, most current research does not fully leverage dermatological knowledge, which can compromise the quality of model predictions. This thesis addresses the ongoing need to link dermatological expertise with machine learning models. To do so, we have closely studied the implementation of a specific dermatological concept known as the "Ugly Duckling" sign. This concept emphasizes comparing lesions within the same patient to improve diagnostic accuracy. The general idea behind this approach is that atypical lesions in the same patient are more likely to be cancerous.

In this thesis, we present various approaches based on this concept, bridging innovations in artificial intelligence with medical advancements.

The first section presents the medical and scientific background of melanoma diagnosis using computer vision.

The second section introduces two approaches we have developed that tackle the task of intra-patient comparison of skin lesions. One is a fully unsupervised approach using visual representation learning and anomaly detection algorithms without annotations. The other is a supervised approach based on neural networks and Transformer architectures.

The third section focuses on unsupervised image encoding through new visual representation learning methods. We present the results of this novel method on melanoma classification.

Finally, the last section addresses model interpretability, particularly through the use of generative models. This approach could eventually enhance the interpretability of artificial intelligence models in clinical settings.

Keywords: Artificial Intelligence, Deep Learning, Visual Representations, Melanoma Diagnosis, Transformers.

Remerciements

Je tiens d'abord à remercier les membres du jury pour l'honneur qu'ils m'ont fait d'examiner ma thèse : Christine Fernandez-Maloigne, Caroline Gaudy, Monique Thonnat et Laurent Heutte.

Je remercie également très chaleureusement mon directeur et ma co-directrice de thèse, Djamal Merad et Séverine Dubuisson, ainsi que mon co-encadrant, Rabah Iguernaissi, qui m'ont permis de réaliser cette thèse et m'ont prodigué de précieux conseils tout au long de ces trois années. Grâce à eux, j'ai énormément appris sur le monde de la recherche.

Je souhaite exprimer ma gratitude aux différents organismes qui ont rendu ce projet possible : le Laboratoire d'Informatique et des Systèmes, l'Université d'Aix-Marseille, l'Agence Nationale de la Recherche pour le projet Diamelex, ainsi que l'École Doctorale de Mathématiques et d'Informatique de Marseille.

Un chaleureux remerciement à tous les membres de l'équipe d'Images et Modèles qui m'ont soutenu tout au long de cette aventure, en me donnant la motivation de continuer et en m'offrant des échanges fructueux. Merci à mes collègues de bureau, Moncef et Adrien, avec qui j'ai partagé de nombreux bons moments. Merci à Arthur, Jilliana et Rabih, avec lesquels j'ai eu la chance de travailler sur des sujets très intéressants et qui m'ont énormément appris dans de nombreux domaines. Merci également aux membres de l'équipe IM : Raoul, Mohammed, Salma, Rim, Sabrine, Mira et tous les autres.

Je tiens à exprimer ma profonde gratitude à ma famille, ma compagne et mes amis, qui, tout au long de ma thèse, m'ont apporté tant de bonheur et de motivation. Je les remercie tous sincèrement pour avoir toujours été là pour moi, pour m'avoir motivé, encouragé et écouté. Merci à ma compagne Chloé, ma mère Anne-Laure, mon frère Geoffrey, Serge ainsi que tous le reste de ma famille, qui m'ont constamment soutenu, et envers qui je ne saurai jamais exprimer pleinement ma reconnaissance.

Je remercie également mes amis proches qui ont rendu cette période plus douce et agréable : Cyril, Pascal, Pierre.

Une pensée particulière à mon chien Billy, que je porterai toujours dans mon cœur, et à mon chat Lucky.

Enfin, je remercie toutes les personnes qui ont été là pour moi durant cette période.

Table des matières

Affidavit 2

Liste de publications et participation aux conférences 3

Résumé 5

Abstract 7

Remerciements 8

Table des matières 9

Table des figures 12

Liste des tableaux 16

Liste des acronymes 17

Glossaire 19

Introduction 20

1. Contexte médical et cadre scientifique 23

1.1. Contexte clinique . . . 23
1.1.1. Le mélanocyte . . . 24
1.1.2. Le mélanome . . . 26
1.1.3. Epidémiologie . . . 29
1.1.4. Prédispositions et facteurs de risque . . . 30
1.1.5. Evolutions du mélanome . . . 31
1.1.6. Prévention . . . 31
1.1.7. Diagnostic . . . 32
1.2. Cadre Méthodologique . . . 36
1.2.1. Bases de données dermatologiques . . . 36
1.2.2. Métriques utilisées . . . 39
1.3. Travaux précédents sur la classification de mélanomes . . . 41
1.3.1. Introduction . . . 41
1.3.2. Les systèmes experts en vision par ordinateur . . . 41
1.3.3. Réseaux de neurones artificiels . . . 42
1.3.4. Applications au mélanome . . . 44
1.3.5. Approches par comparaison d'images . . . 46

1.4. Synthèse 47

2. Comparaison des lésions intra-patient pour la détection du mélanome 49
2.1. Introduction 50
2.2. Revue de la littérature 50
2.2.1. Caractérisation du VPC par des participants 51
2.2.2. Analyse de la perte d'AutoEncodeurs 51
2.2.3. Calcul des distances dans l'espace latent 53
2.2.4. Approches basées sur l'attention 54
2.2.5. Méthodes de détection d'anomalies 55
2.3. Détection non supervisée des mélanomes 55
2.3.1. Introduction 55
2.3.2. Méthode 56
2.3.3. Résultats 60
2.3.4. Conclusion 62
2.4. Détection supervisée du mélanome par comparaison de lésions cutanées par patient 63
2.4.1. Introduction 63
2.4.2. Méthode 64
2.4.3. Résultats 67
2.4.4. Discussion 70
2.5. Synthèse 70

3. Apprentissage auto-supervisé de représentations visuelles 72
3.1. Introduction 72
3.2. Etat de l'art 74
3.2.1. Origines du domaine 74
3.2.2. Méthodes actuelles 75
3.2.3. Limites et perspectives 80
3.3. Conception d'un nouveau modèle d'apprentissage de représentations visuelles 81
3.3.1. Architecture proposée 81
3.3.2. Expériences sur Caltech-256 84
3.3.3. Expériences sur les lésions cutanées 90
3.3.4. Discussion 91
3.4. Synthèse 92

4. La génération d'images, une voie vers l'interprétabilité 94
4.1. Introduction 94
4.2. Etat de l'art 95
4.2.1. Approches générales d'interprétabilité 96
4.2.2. La génération d'images de manière générale 99
4.2.3. L'interprétabilité à travers la génération d'images 102

4.3. Interprétation de classifieurs par génération d'images 103
4.3.1. La génération d'images difficiles à diagnostiquer 104
4.3.2. Méthodologie 104
4.3.3. Gestion des artefacts 105
4.3.4. Images générées 105
4.3.5. Impact de l'ajout des images synthétiques sur la classification . 108
4.3.6. Analyse de l'espace latent 109
4.4. Synthèse 110

Conclusion **112**

Perspectives **114**

Bibliographie **116**

ANNEXES **130**
A. Travaux sur l'asymétrie des lésions cutanées 130
B. Travaux sur la théorie des jeux appliquée à la détection de maladies cutanées 130
C. Travaux sur la notion de désordre en dermatologie 130

Table des figures

1.1. **Exemple de mélanocyte sain.** Un mélanocyte seul (en violet) et dans le tissu cutané. *Image par Blausen Medical[2]* 25
1.2. **Représentation schématique du mélanome se développant sous la peau.** Les différents stades du développement du mélanome sont présentés, de Tis (in situ, i.e présent seulement à la couche supérieure de la peau) à T4 (le mélanome est long de 4mm ou plus). *Image par Cancer Research UK, CC BY-SA 4.0* 26
1.3. **Un dermatoscope (Heine, delta-10).** C'est un outil très utilisé en clinique, permettant une bonne illumination de la lésion lors de l'analyse visuelle. Il est également possible de l'utiliser en complément d'un appareil photo pour prendre des images détaillées des lésions. *Image par Frank33, sous license CC BY-SA 3.0 et GNU FDL.* 27
1.4. **Image dermoscopique d'un mélanome.** On remarque plusieurs caractéristiques typiques du mélanome, avec une asymmétrie présente, des bords irréguliers ainsi que le présence de plusieurs couleurs. *Image d'ISIC 2020 [3]* 28
1.5. **Exemple de dix dermoscopies.** *(en haut)* cinq mélanomes *(en bas)* cinq naevus. *Images d'ISIC 2020 [3]* 28
1.6. **Décès dus au mélanome et autres cancers de la peau en 2012 par million de personnes.** *Statistiques de l'OMS, regroupées par déciles. Carte par BlankMap-World6.* 29
1.7. **Schéma représentant les différents critères de la règle ABCDE.** 33
1.8. **Schéma représentant le concept du Vilain Petit Canard pour le diagnostic du mélanome.** Intuitivement, cela correspond à la détection d'une lésions très différente des autres. Informatiquement, cela se traduira par une grande distance dans l'espace de représentations. 35
1.9. **Schéma simplifié d'un réseau de neurones à quatre couches.** 43
1.10. **Évolution des performances de classification d'images au fil des années sur la base de données ImageNet [30].** En 2013, le réseau de neurones convolutif AlexNet [31] a grandement amélioré les performances. *Crédit image : [32]* 44

2.1. **Exemple de résultat de l'approche utilisant un AutoEncodeur.** Les lésions détectées sont ensuite ordonnées par la perte qu'elles ont obtenues en passant par l'AutoEncodeur entraîné sur les lésions cutanées. *Image de l'article original [57].* 52

2.2. **Présentation de l'architecture de l'article [60] basé sur la distance au centroïde.** Un CNN de classification est entraîné à détecter les mélanomes des naevus, puis les plongements obtenus sont comparés au centroïde du nuage de point obtenu par patient. *Image provenant de l'article original [60]* . . . 53
2.3. **Architecture de notre approche**. L'encodeur transforme les images en vecteurs, puis ces derniers sont comparés à l'aide de l'algorithme des k-plus-proches-voisins. Enfin, le seuil optimal est choisi empiriquement sur la base de validation. . . . 57
2.4. **Diagramme de l'architecture SimSiam [72].** *Image provenant de l'article original.* . . . 57
2.5. **Exemples de la data augmentation utilisée.** L'image originale est affichée tout à gauche, suivie de la même image transformée aléatoirement de 4 manière différentes. . . . 59
2.6. **Visualisation de notre algorithme de détection d'anomalies.** Dans cet exemple, $k = 1$ est utilisé sur un nuage de points en 2 dimensions. Chaque point représente le vecteur de caractéristiques d'une lésion tandis que sa couleur représente son UDI. . . . 60
2.7. **Performance sur la détection de mélanome en fonction de k. k=0 compare les lésions avec le centroid.** . . . 61
2.8. **Résultats sur la base de test après avoir choisi k=3 grâce à la base de validation.** . . . 62
2.9. **Diagramme de l'architecture ReSeT**. L'encodeur prend en entrée les images une par une, puis le Set-Transformer récupère tous les vecteurs d'un coup pour les comparer. Enfin, les sorties du Set-Transformer sont concaténées avec les vecteurs issus de l'encodeur afin de générer la prédiction finale bénin/malin. . . . 65
2.10. **Diagramme de l'architecture des Set-Transformers et ses différentes opérations.** . . . 66
2.11. **Courbe ROC avec écart-type sur la détection de mélanomes en fonction de chaque encodeur.** . . . 69

3.1. **Schéma représentant l'architecture MoCo.** L'encodeur principal extrait des caractéristiques des images d'entrée et est mis à jour à chaque itération du processus d'entraînement, tandis que le momentum encoder est une version stabilisée de l'encodeur principal, qui est mise à jour lentement à l'aide d'un coefficient de momentum. *Image provenant de l'article original [78].* . . . 76
3.2. **Schéma représentant l'architecture Byol.** Chacune des branches reçoit une vue augmentée d'une même image d'entrée. Le premier modèle, appelé "target network", est mis à jour de manière lente et accumulative, ce qui lui permet d'agir comme une référence stable. Le second modèle, ou "online network", est mis à jour de manière plus rapide à chaque itération. . . . 77

3.3. **Exemple d'images reconstruites par le modèle MAE.** On constate que le modèle reste capable de reconstruire une bonne partie de l'image malgré une grande proportion de patches masqués. *Image provenant de l'article initial [86]* 79
3.4. **Vue d'ensemble de l'architecture SJ-MAE.** Un MAE classique est implémenté avec ses modules encodeur-décodeur. Nous ajoutons deux nouvelles branches correspondant à la tâche de puzzle et à la tâche siamoise. Le module puzzle prend les plongements extraits de l'encodeur, sans embeddings positionnels, et résout le puzzle sur les patches visibles, tandis que la branche siamoise rapproche les deux plongements l'un de l'autre. 81
3.5. **Exemple d'images de Caltech-256.** 84
3.6. **Valeurs des fonctions de pertes une fois équilibrées.** 86
3.7. **Précision des différents modèles par nombre d'époques de pré-entraînement.** 88
3.8. **Comparaison des résultats de la classification linéaire et du fine-tuning en fonction de différents taux de masquage.** Résultats obtenus après 100 époques de pré-entraînement, suivies de 50 époques de fine-tuning ou de 100 époques de classification linéaire. 89
3.9. **Exemples de résultats sur les images de validation de Caltech-256.** (à gauche) l'image masquée, (au milieu) la reconstruction par SJ-MAE, (à droite) vérité terrain. Le ratio de masquage est de 80 %, ne laissant que 39 patchs sur 196. 90
3.10. **Exemples de résultats sur les images de validation d'ISIC 2020.** (à gauche) vérité terrain, (au milieu) l'image masquée, (à droite) reconstruction des patchs manquants par SJ-MAE. Le ratio de masquage est de 50 %, ne laissant que 98 patchs sur 196. 92

4.1. **Exemple de carte de chaleur obtenue avec Grad-CAM.** Ici, les zones rouges représentent les endroits où le modèle a un gradient élevé pour la catégorie "chat". 96
4.2. **Quatre exemples de cartes de chaleur sur des images dermoscopiques.** Le rouge représente les zones à forte activation de gradient pour la catégorie 'mélanome'. On remarque que les cartes de chaleur se focalisent effectivement sur les lésions, ce qui montre que le modèle se concentre au bon endroit pour réaliser le diagnostic. Toutefois, cette information n'est pas directement utile aux dermatologues lors de l'analyse des lésions. 97
4.3. **Visualisation de filtres appris par un CNN.** On constate que le modèle a appris à reconnaître des formes complexes dans le cas d'images de chiens, par exemple. *Image provenant de [107]* 98
4.4. **Images générées par le modèle avec la condition de générer un mélanome bien prédit par le classifieur.** 106
4.5. **Images générées par le modèle avec la condition de générer un mélanome mal prédit par le classifieur.** 106

4.6. **Images générées par le modèle avec la condition de générer un naevus bien prédit par le classifieur.** 107
4.7. **Images générées par le modèle avec la condition de générer un nevus mal prédit par le classifieur.** 107
4.8. **Projection des images d'ISIC 2020 en deux dimensions à l'aide des vecteurs appris par classification. Les points jaunes sont des naevus et les points violets des mélanomes. Il est possible de visualiser l'image choisie en bas de l'écran.** Nous remarquons différents clusters, présentant chacun des caractéristiques visuelles propres. 109

Liste des tableaux

1.1. **Présentation des principales bases de données d'images dermoscopiques, affichées par nombre croissant d'images.** [1][22] [2][23] [3][14] [4][15] [5][16] [6][17, 18] [7][19-21] [8][3] 37

2.1. **AUC et précision équilibrée de la classification sur la base de test.** Nous comparons les performances entre un CNN de classification classique, un SetTransformer entrainé sur les caractéristiques du CNN ainsi que les performances de ReSeT sur les caractéristiques du CNN. 68
2.2. **Sensitivité et spécificité de la classification sur la base de test.** 68
2.3. Résultats de la tâche de classification sur le jeu de test utilisant des architectures auto-supervisées pour l'entraînement du CNN. 68
2.4. **Résultats de la tâche de classification sur le jeu de test utilisant des architectures auto-supervisées pour l'entraînement du CNN.** 69

3.1. **Valeurs des différentes pertes en fonction du nombre d'étapes.** En moyenne, la perte du puzzle est 7 fois plus élevée que la perte de reconstruction. 85
3.2. **Précision des variantes MAE.** Les meilleures précision par colonne sont affichée en gras. 87
3.3. **Précision des variantes MAE.** Les meilleures précision par colonne sont affichée en gras. 91
3.4. **Résultats de la tâche de classification sur le jeu de test utilisant les architectures auto-supervisées pour l'entraînement du CNN.** 91

4.1. **Comparaison des performances lors de l'ajout d'un poids de classe et lors de l'utilisation de données synthétiques.** Les meilleures valeurs par métrique sont présentées en gras pour chaque ligne. L'utilisation de données synthétiques montre une performance supérieure. 108

Liste des acronymes

AE
Autoencodeur –ou *Autoencoder*–.

AI
Intelligence Artificielle –ou *Artificial Intelligence*–.

AUC
Area Under the Curve, une métrique qui mesure l'aire sous la courbe ROC pour évaluer la capacité globale d'un classifieur binaire à distinguer entre les classes.

BCE
Binary Cross-Entropy, une fonction de perte utilisée pour les modèles de classification binaire.

CNN
Réseau de Neurones Convolutifs –ou *Convolutional Neural Network*–.

DL
Apprentissage Profond –ou *Deep Learning*–.

FNR
Taux de Faux Négatifs –ou *False Negative Rate*–.

FPR
Taux de Faux Positifs –ou *False Positive Rate*–.

GAN
Réseau Antagoniste Génératif –ou *Generative Adversarial Network*–.

ISIC
International Skin Imaging Collaboration, une base de données standardisée pour la détection des lésions cutanées, incluant des images de mélanomes.

LIME
Local Interpretable Model-agnostic Explanations, une méthode d'explicabilité des modèles de machine learning.

ML
Apprentissage Automatique –ou *Machine Learning*–.

RL

Apprentissage de Représentations –ou *Representation Learning*–.

ROC

Receiver Operating Characteristic, une courbe qui illustre les performances d'un classifieur binaire en traçant le taux de vrais positifs en fonction du taux de faux positifs pour différents seuils de décision.

SHAP

SHapley Additive exPlanations, une méthode d'explicabilité basée sur la théorie des jeux.

SSL

Apprentissage Auto-Supervisé –ou *Self-Supervised Learning*–.

UL

Apprentissage Non Supervisé –ou *Unsupervised Learning*–.

ViT

Transformeur Visuel –ou *Vision Transformer*–.

Glossaire

Apprentissage Auto-Supervisé

Méthode d'apprentissage où le modèle utilise des données non étiquetées et génère ses propres labels à partir de signaux inhérents aux données. Elle est utilisée dans des domaines comme la vision par ordinateur, où des annotations manuelles peuvent être coûteuses..

Apprentissage de Représentations

Sous-domaine de l'apprentissage automatique où l'objectif est d'apprendre des caractéristiques utiles et compactes à partir de données brutes, souvent utilisé pour améliorer les performances dans des tâches comme la classification d'images..

Autoencodeur

Un autoencodeur est un réseau neuronal non supervisé conçu pour apprendre une représentation compacte des données en reconstruisant son entrée. Il est souvent utilisé pour la réduction de dimension et la détection d'anomalies..

Plongement

Représentation vectorielle de données dans un espace continu de faible dimension, utilisée pour capturer les relations sémantiques ou structurelles entre les entités représentées. Aussi appelé *embedding* en anglais, le plongement est largement employé en apprentissage automatique, notamment pour des applications telles que le traitement du langage naturel ou la vision par ordinateur.

Introduction

Ces dernières années, la prévalence du mélanome n'a cessé d'augmenter dans les pays développés [1]. Bien que ce cancer de la peau soit l'un des plus rares, il demeure l'un des plus mortels en raison de sa capacité à métastaser rapidement. Dans ce contexte, il est essentiel d'améliorer les différentes étapes de la prise en charge, du diagnostic à la guérison.

Dans cette thèse, nous nous sommes concentrés sur la première étape de la prise en charge, à savoir la détection du mélanome. À un stade précoce de la maladie, une simple exérèse locale permet de guérir le patient, tandis qu'une fois le mélanome développé, des traitements lourds et longs sont nécessaires, avec des chances de survie moindres. Il est donc crucial de détecter le mélanome le plus tôt possible. Cependant, le diagnostic à ce stade est difficile, car les mélanomes peuvent ressembler aux naevus, aussi appelés grains de beauté, qui sont des lésions bénignes de la peau, souvent nombreuses, et qu'il est inutile de retirer.

C'est dans ce contexte, que l'ANR a financé un projet collaboratif DIAMELEX (https ://anr.fr/Projet-ANR-20-CE45-0026) pour améliorer le diagnostic de ce cancer, en finançant plusieurs thèses. La mienne s'est particulièrement focalisée sur le concept du « vilain petit canard » et ses applications pratiques. Ce concept, encore peu exploité dans la littérature, sera présenté et développé tout au long de la thèse. L'idée générale est qu'une lésion différente des autres chez un même patient a un risque plus élevé d'être cancéreuse. L'utilisation de cette information pourrait grandement améliorer les chances de détection du cancer et donc augmenter la survie des patients. Cela soulève des problématiques à la fois de comparaison et de compréhension d'images par des algorithmes automatiques, que nous développerons au fil des chapitres.

Ainsi, la collaboration étroite entre les différents acteurs de ce projet a conduit à la publication de plusieurs avancées notables, nourries par la diversité des idées et des perspectives issues de l'hôpital de la Timone, de l'entreprise Anapix Medical, ainsi que des chercheurs du Cnam et du Laboratoire d'Informatique et Systèmes, dont j'ai fait partie. Nous espérons que ces publications auront un impact sur le domaine médical et contribueront, à terme, au développement de nouvelles technologies facilitant et accélérant la prise en charge des patients atteints de maladies dermatologiques.

Sur un plan plus personnel, ma thèse s'inscrit dans la continuité des travaux que j'ai réalisés lors de mon stage l'année précédant ma thèse, sur l'importance de l'asymétrie dans la détection du mélanome. Ces travaux ont également abouti à la publication d'un article dans une revue scientifique. Durant ma thèse, j'ai aussi eu l'opportunité de collaborer sur d'autres problématiques liées à mon sujet, en accord avec d'autres doctorants et chercheurs du LIS, tels qu'Arthur Foahom Gouabou et Jilliana Monnier. Ces collaborations ont permis la publication de plusieurs articles de conférence dont

je suis co-auteur. Bien que ces travaux ne soient pas détaillés dans cette thèse, ils peuvent être consultés dans les thèses respectives des collaborateurs mentionnés ci-dessus.

Bien que centrées sur le mélanome, les approches développées peuvent également être adaptées à d'autres domaines, ce qui pourrait inspirer les lecteurs travaillant dans des disciplines confrontées à des problématiques similaires à celles du mélanome.

En pratique, nous espérons que les approches présentées auront un impact clinique, en offrant aux praticiens un outil d'assistance automatisé, tel un second avis lors du diagnostic, et des informations diverses sur les lésions et leur ressemblance ou dissemblance avec d'autres lésions présentes chez le patient, ce qui permettrait d'améliorer le diagnostic.

Enfin, nous espérons que les travaux présentés ici contribueront à faire progresser la recherche dans les domaines médicaux et informatiques, et qu'ils encourageront le développement de nouveaux projets alliant informatique et médecine, dans le but d'aider davantage la population et de lutter contre des maladies qui, trop souvent encore, affectent gravement nos vies.

En ce qui concerne l'organisation de cette thèse, elle est structurée en quatre chapitres :

Le Chapitre 1 présente le contexte médical et scientifique du diagnostic du mélanome, en détaillant l'intérêt de notre approche. Dans le contexte médical, nous abordons les caractéristiques physiques des mélanomes, l'épidémiologie et les différentes causes possibles, puis nous décrivons les méthodes utilisées par les dermatologues pour détecter ce cancer. Sur le plan scientifique, nous résumons les avancées récentes en matière d'intelligence artificielle qui permettent de classer et comparer les images. Enfin, nous exposons le cadre méthodologique actuel, notamment les bases de données et les métriques utilisées.

Le Chapitre 2 présente les approches que nous avons développées pour la comparaison et la classification des mélanomes, à la fois de manière auto-supervisée et supervisée. Nous montrons que les approches auto-supervisées peuvent offrir de bonnes performances tout en s'affranchissant de la nécessité d'annotations, puis nous concevons un nouveau modèle de comparaison d'images de mélanomes, à la fois simple, rapide et efficace.

Le Chapitre 3 aborde une problématique plus générale, celle de la capacité des réseaux de neurones à générer des représentations visuelles à partir d'images sans l'aide d'annotations. Ces modèles auto-supervisés pourraient améliorer les tâches de comparaison et de classification en générant de meilleures représentations visuelles. Les méthodes présentées sont utiles non seulement pour la détection du mélanome, mais aussi pour d'autres tâches similaires.

Le Chapitre 4 expose nos travaux les plus récents ainsi que les perspectives dans le domaine de l'interprétabilité des réseaux de neurones appliqués aux lésions cutanées. Nous rappelons les différentes méthodes d'interprétabilité existantes et mettons en avant celles qui sont les plus pertinentes pour notre problématique. Enfin, nous présentons nos derniers travaux et les perspectives futures dans ce domaine.

1. Contexte médical et cadre scientifique

Sommaire

1.1. Contexte clinique 23
1.1.1. Le mélanocyte 24
1.1.2. Le mélanome 26
1.1.3. Epidémiologie 29
1.1.3.1. Incidence 29
1.1.3.2. Mortalité 29
1.1.4. Prédispositions et facteurs de risque 30
1.1.4.1. Facteurs endogènes 30
1.1.4.2. Facteurs exogènes 30
1.1.5. Evolutions du mélanome 31
1.1.5.1. Formes superficielles 31
1.1.5.2. Mélanome nodulaire 31
1.1.6. Prévention 31
1.1.7. Diagnostic 32
1.1.7.1. Règle "ABCDE" 33
1.1.7.2. Le concept du “Vilain Petit Canard” (VPC) 35
1.1.7.3. Les informations patient 36
1.2. Cadre Méthodologique 36
1.2.1. Bases de données dermatologiques 36
1.2.2. Métriques utilisées 39
1.3. Travaux précédents sur la classification de mélanomes 41
1.3.1. Introduction 41
1.3.2. Les systèmes experts en vision par ordinateur 41
1.3.3. Réseaux de neurones artificiels 42
1.3.4. Applications au mélanome 44
1.3.5. Approches par comparaison d’images 46
1.4. Synthèse 47

1.1. Contexte clinique

Tout d’abord, nous allons présenter le contexte clinique du mélanome et de sa détection. Dans cette partie, nous décrirons en détail les cellules initiales que sont

les mélanocytes, ainsi que leur fonctionnement et utilité pour le corps humain. Nous verrons comment, à partir de ces cellules, des cancers peuvent se former et poser des problèmes aux patients. Nous étudierons également l'épidémiologie à travers le monde, les facteurs de risque, puis nous présenterons les différentes approches utilisées en clinique pour les détecter et soigner les patients.

1.1.1. Le mélanocyte

Les mélanocytes (voir Figure 1.1) sont des cellules spécialisées présentes dans l'épiderme de la peau, ainsi que dans d'autres tissus comme les cheveux, les yeux, et même les muqueuses. Leur fonction principale est la production de mélanine, un pigment responsable de la coloration de la peau, des cheveux et des yeux, ainsi que de la protection contre les effets nocifs des rayons ultraviolets (UV) du soleil.

Au niveau de l'épiderme, les mélanocytes sont localisés dans la couche basale, la couche la plus profonde de l'épiderme, où ils interagissent étroitement avec les kératinocytes, les cellules majoritaires de l'épiderme. Les mélanocytes ont une forme étoilée avec de longues extensions appelées dendrites qui s'étendent entre les kératinocytes.

Le processus de production de mélanine, la mélanogenèse, se déroule dans des organites spécialisés appelés mélanosomes, qui sont des vésicules intracellulaires. Les mélanocytes synthétisent la mélanine à partir de l'acide aminé tyrosine, via une série de réactions enzymatiques complexes impliquant des enzymes telles que la tyrosinase.

Une fois formée, la mélanine est transférée des mélanocytes aux kératinocytes avoisinants par le biais des dendrites des mélanocytes. Cette mélanine agit comme un écran solaire naturel en absorbant les rayons UV du soleil, réduisant ainsi les dommages causés à l'ADN des cellules cutanées et protégeant contre les dommages causés par les UV tels que les coups de soleil et le cancer de la peau.

Outre leur rôle dans la pigmentation de la peau, les mélanocytes jouent également un rôle dans d'autres processus biologiques, comme leur implication dans la réponse immunitaire de la peau et peuvent agir comme des cellules sentinelles en détectant les dommages causés par les UV et en déclenchant des réponses inflammatoires pour protéger la peau.

Malheureusement, des anomalies dans les mélanocytes peuvent entraîner des maladies dermatologiques telles que le vitiligo (une maladie auto-immune caractérisée par la perte de pigmentation de la peau) ou encore le mélanome, qui est un type de cancer de la peau dérivé des mélanocytes et que nous étudierons tout au long de cette thèse.

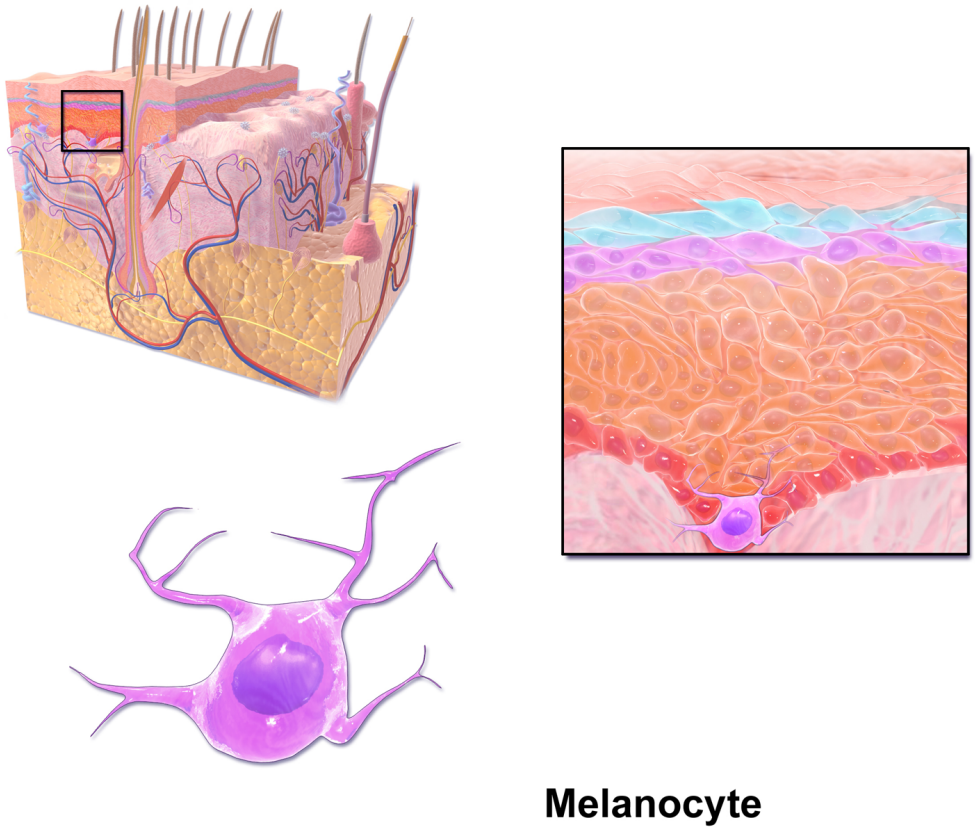

FIGURE 1.1. – **Exemple de mélanocyte sain.** Un mélanocyte seul (en violet) et dans le tissu cutané. *Image par Blausen Medical[2]*

1.1.2. Le mélanome

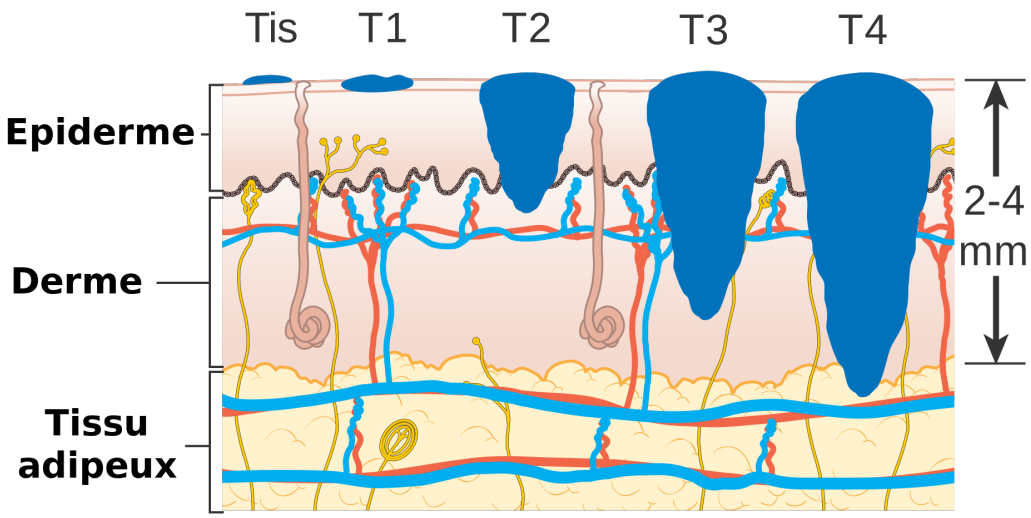

FIGURE 1.2. – **Représentation schématique du mélanome se développant sous la peau.** Les différents stades du développement du mélanome sont présentés, de Tis (in situ, i.e présent seulement à la couche supérieure de la peau) à T4 (le mélanome est long de 4mm ou plus). *Image par Cancer Research UK, CC BY-SA 4.0*

Le mélanome peut survenir de manière sporadique ou être associé à des facteurs de risque tels que l'exposition excessive aux rayons UV du soleil, les antécédents familiaux de mélanome, les coups de soleil sévères pendant l'enfance, ainsi que des conditions génétiques prédisposantes. Les mutations génétiques, en particulier celles affectant les gènes impliqués dans la régulation de la croissance cellulaire et la réparation de l'ADN, jouent un rôle clé dans le développement du mélanome.

Cliniquement, le mélanome se manifeste souvent par l'apparition de nouvelles lésions cutanées ou par des changements dans les lésions existantes, notamment des changements de taille, de forme, de couleur ou de texture. Il peut apparaître sous différentes formes, y compris des taches pigmentées, des grains de beauté, des nodules ou des plaques irrégulières.

Pour le détecter, la première étape consiste à analyser la peau, généralement réalisée en clinique à l'aide d'un dermatoscope (voir Figure 1.3), permettant de prendre des images dermoscopiques. Ce type particulier d'image permet une bonne illumination des lésions, ce qui permet de voir les détails intrinsèques de celles-ci. Les Figures 1.4 et 1.5 présentent quelques exemples d'images dermoscopiques issues de bases de données publiques que nous décrirons dans les parties suivantes. La majorité des travaux d'aide au diagnostic sont réalisés sur ce type d'images, et la totalité de nos

expériences au sein de cette thèse sont effectuées sur ce type de données.



FIGURE 1.3. – **Un dermatoscope (Heine, delta-10).** C'est un outil très utilisé en clinique, permettant une bonne illumination de la lésion lors de l'analyse visuelle. Il est également possible de l'utiliser en complément d'un appareil photo pour prendre des images détaillées des lésions. *Image par Frank33, sous license CC BY-SA 3.0 et GNU FDL.*

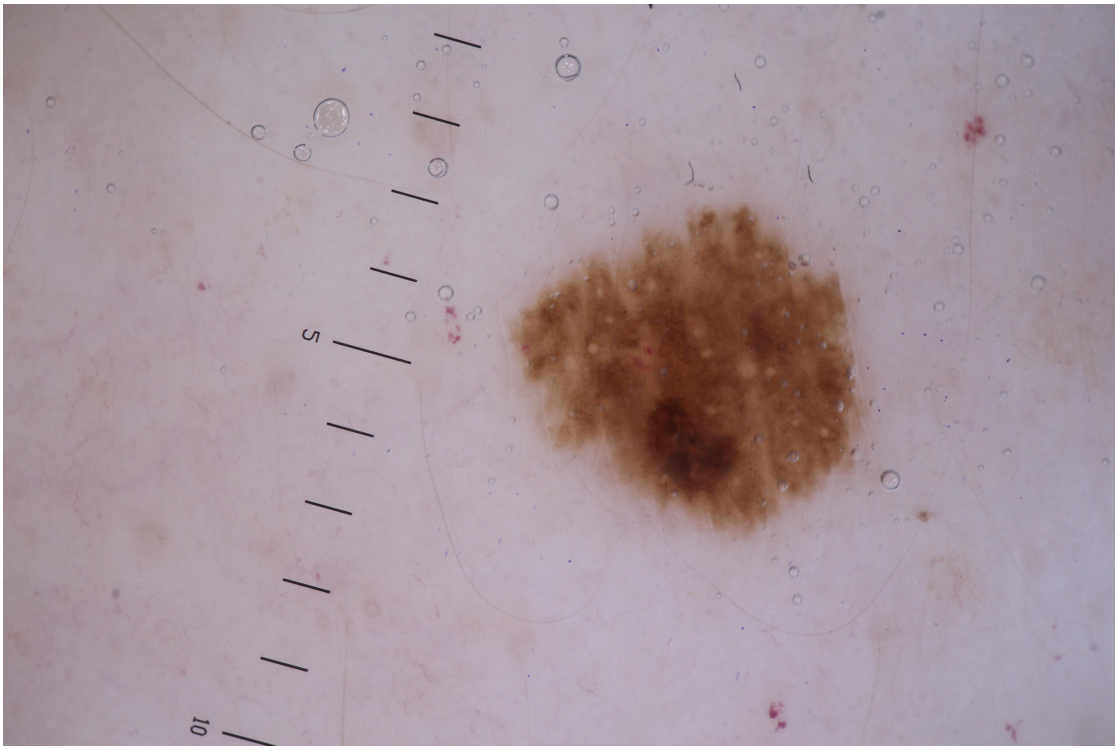

FIGURE 1.4. – **Image dermoscopique d'un mélanome.** On remarque plusieurs caractéristiques typiques du mélanome, avec une asymmétrie présente, des bords irréguliers ainsi que le présence de plusieurs couleurs. *Image d'ISIC 2020 [3]*

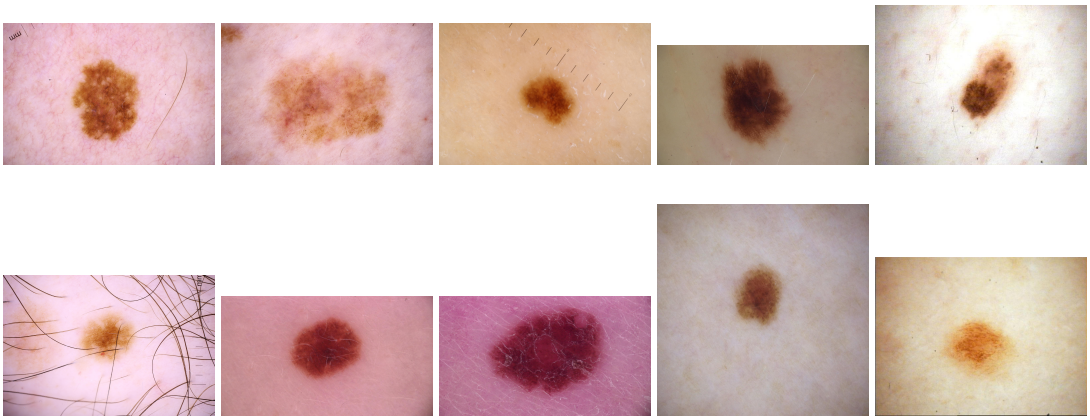

FIGURE 1.5. – **Exemple de dix dermoscopies.** *(en haut)* cinq mélanomes *(en bas)* cinq naevus. *Images d'ISIC 2020 [3]*

1.1.3. Epidémiologie

Le mélanome cutané représente un problème croissant de santé publique dans de nombreuses régions du monde, avec des taux d'incidence en constante augmentation au cours des dernières décennies [4], [5], [6]. L'épidémiologie du mélanome comprend deux aspects principaux : l'incidence, qui fait référence au nombre de nouveaux cas diagnostiqués chaque année, et la mortalité, qui se réfère au nombre de décès dus au mélanome.

1.1.3.1. Incidence

L'incidence du mélanome varie considérablement selon la région géographique, la couleur de peau, le sexe et l'âge. Généralement, les régions avec une forte exposition aux rayons UV du soleil, telles que l'Australie, la Nouvelle-Zélande et les régions d'Europe occidentale et du nord de l'Amérique, ont les taux d'incidence les plus élevés. Cependant, le mélanome peut également survenir dans des régions moins ensoleillées et chez des personnes à la peau plus sombre, bien que moins fréquemment.

L'incidence du mélanome a augmenté de manière significative au cours des dernières décennies. Cette augmentation est attribuée à plusieurs facteurs, notamment l'exposition accrue aux rayons UV du soleil due aux comportements de bronzage, à l'amélioration de la sensibilisation et du dépistage, ainsi qu'à des facteurs environnementaux et génétiques.

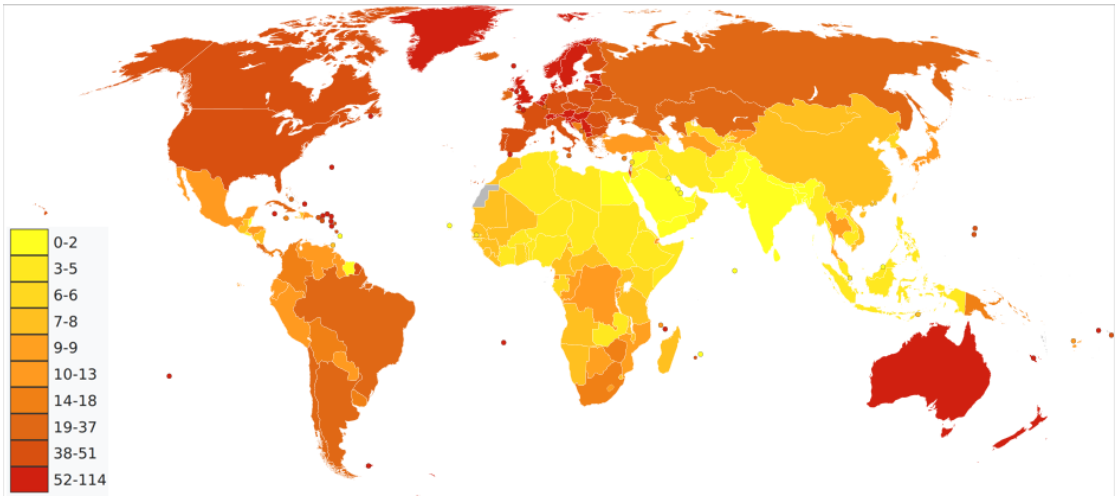

FIGURE 1.6. – **Décès dus au mélanome et autres cancers de la peau en 2012 par million de personnes.** *Statistiques de l'OMS, regroupées par déciles. Carte par BlankMap-World6.*

1.1.3.2. Mortalité

Bien que le mélanome soit généralement associé à un bon pronostic lorsqu'il est diagnostiqué tôt, il peut devenir mortel s'il n'est pas traité à temps ou s'il se propage

à d'autres parties du corps. La mortalité due au mélanome dépend principalement du stade de la maladie au moment du diagnostic. Les mélanomes diagnostiqués à un stade précoce, lorsque la tumeur est confinée à la peau, ont un pronostic favorable avec des taux de survie élevés. En revanche, les mélanomes avancés, qui se sont propagés à d'autres organes ou tissus (métastases), ont un pronostic moins favorable et peuvent être difficiles à traiter.

Bien que la mortalité globale due au mélanome reste relativement faible par rapport à d'autres types de cancer, elle peut être considérable dans les cas avancés. Les progrès dans le dépistage précoce, le traitement et la sensibilisation sont essentiels pour réduire la mortalité liée au mélanome et améliorer les résultats cliniques pour les patients atteints de cette maladie.

1.1.4. Prédispositions et facteurs de risque

Le développement du mélanome est influencé par une combinaison complexe de prédispositions génétiques et de facteurs environnementaux. Ces facteurs peuvent être classés en deux catégories principales : les facteurs endogènes, qui sont inhérents à l'individu, et les facteurs exogènes, qui proviennent de l'environnement externe.

1.1.4.1. Facteurs endogènes

Les facteurs endogènes comprennent les caractéristiques génétiques et biologiques de l'individu. Les antécédents familiaux de mélanome jouent un rôle significatif, car les individus ayant des membres de la famille proche diagnostiqués avec un mélanome ont un risque accru de développer la maladie [7]. Les mutations génétiques héritées, telles que celles dans les gènes CDKN2A ou CDK4, peuvent également prédisposer à un risque accru de mélanome [8]. De plus, le type de peau, avec une prévalence plus élevée chez les personnes à la peau claire, aux cheveux blonds ou roux, et aux yeux clairs, ainsi que la présence de grains de beautés atypiques ou de naevus congénitaux, peuvent augmenter le risque.

1.1.4.2. Facteurs exogènes

Les facteurs exogènes, quant à eux, sont des éléments externes à l'individu. L'exposition aux rayons ultraviolets (UV) du soleil est le principal facteur de risque environnemental du mélanome. Une exposition excessive aux UV, notamment les coups de soleil sévères pendant l'enfance ou l'adolescence, ainsi que l'utilisation de lits de bronzage et d'autres sources artificielles de lumière UV, accroît considérablement le risque [9]. De plus, certains métiers impliquant une exposition prolongée au soleil, comme les travailleurs de la construction ou les agriculteurs, présentent un risque accru de mélanome. Enfin, l'immunosuppression, qu'elle soit due à des médicaments immunosuppresseurs après une transplantation d'organe ou à des maladies auto-immunes telles que le VIH/SIDA, peut également augmenter le risque de développer un mélanome. [10]

1.1.5. Evolutions du mélanome

Une fois présent, le mélanome peut se développer sous plusieurs formes. D'une part, il existe des formes superficielles, qui se développent principalement à la surface de la peau et adoptent une croissance horizontale pouvant perdurer avant de devenir invasives. D'autre part, les mélanomes nodulaires se caractérisent par une progression rapide, avec une infiltration sous-cutanée dès les premiers stades.

1.1.5.1. Formes superficielles

Les formes superficielles de mélanome représentent un sous-type courant de cette maladie cutanée. Elles sont souvent caractérisées par une croissance horizontale à la surface de la peau, plutôt que par une infiltration verticale. Les formes superficielles incluent le mélanome lentigo maligna, le mélanome acral lentigineux, et le mélanome de type lentigo.

Le mélanome lentigo maligna se développe généralement sur des zones de peau exposées au soleil, telles que le visage et les bras. Il se présente souvent comme une tache irrégulière de couleur brune ou noire, qui peut changer de forme et de taille au fil du temps.

Le mélanome acral lentigineux se forme sur les paumes des mains, les plantes des pieds ou sous les ongles. Contrairement à d'autres formes de mélanome, il n'est pas toujours associé à une exposition au soleil et peut être plus difficile à diagnostiquer.

Le mélanome de type lentigo se développe sur les zones de peau endommagées par le soleil, telles que le visage et les bras. Il apparaît généralement comme une tache plate et irrégulière de couleur brune ou noire, avec des bords mal définis.

1.1.5.2. Mélanome nodulaire

Le mélanome nodulaire est un type agressif de cancer de la peau qui se développe rapidement et se présente souvent comme un nodule élevé sur la peau. Contrairement aux formes superficielles, le mélanome nodulaire a tendance à pousser en profondeur dans la peau dès le début, plutôt que de se propager horizontalement. Cela le rend souvent plus difficile à détecter à un stade précoce.

Les nodules de mélanome peuvent varier en taille, en couleur et en texture, mais ils sont généralement plus foncés que la peau environnante et peuvent saigner ou démanger. En raison de sa croissance rapide et de son potentiel de propagation métastatique, le mélanome nodulaire est souvent considéré comme un type plus grave de cancer de la peau et nécessite une intervention médicale rapide.

1.1.6. Prévention

Ainsi, la prévention est d'une importance capitale pour réduire l'incidence de ce cancer. Alors que le diagnostic précoce est crucial pour un traitement efficace, adopter des mesures préventives peut contribuer à réduire le risque de développement du mélanome. Les différentes formes de préventions incluent :

Protection solaire L'une des mesures préventives les plus efficaces contre le mélanome est la protection solaire adéquate. L'exposition aux rayons ultraviolets (UV) du soleil est l'un des principaux facteurs de risque de développement du mélanome. Il est donc essentiel d'adopter des pratiques de protection solaire telles que l'utilisation de crème solaire à large spectre avec un FPS élevé, le port de vêtements protecteurs, tels que des chapeaux à larges bords et des vêtements à manches longues, et la recherche d'ombre pendant les heures de pointe d'ensoleillement.

Éducation et sensibilisation L'éducation et la sensibilisation jouent un rôle crucial dans la prévention du mélanome. Il est important d'informer le public sur les dangers de l'exposition aux UV et sur les mesures qu'ils peuvent prendre pour réduire leur risque de mélanome. Cela peut inclure des campagnes de sensibilisation dans les écoles, les lieux de travail et les communautés locales, ainsi que des programmes éducatifs sur la protection solaire et la détection précoce du mélanome.

Surveillance des changements cutanés Une surveillance régulière des changements cutanés est une autre composante importante de la prévention du mélanome. Les individus devraient être encouragés à examiner régulièrement leur peau à la recherche de nouvelles lésions ou de changements dans les lésions existantes, en utilisant la règle ABCDE comme guide. Tout changement suspect devrait être signalé à un professionnel de la santé pour une évaluation plus approfondie.

Éviter les lits de bronzage L'utilisation de lits de bronzage est fortement associée à un risque accru de mélanome. Il est donc recommandé d'éviter les lits de bronzage et de privilégier les méthodes de bronzage sans danger telles que les lotions autobronzantes ou les sprays bronzants.

Prévention primaire Enfin, la prévention primaire du mélanome implique d'éviter les facteurs de risque connus tels que les coups de soleil graves, l'exposition aux UV artificiels, les antécédents familiaux de mélanome et les antécédents personnels de lésions cutanées précancéreuses. En adoptant un mode de vie sain et en prenant des mesures préventives appropriées, il est possible de réduire le risque de mélanome et de protéger la santé de la peau à long terme.

1.1.7. Diagnostic

Les patients sont alors fortement encouragés à faire vérifier leur peau chez les dermatologues. Comme vu précédemment, le diagnostic est effectué en clinique lors d'une analyse visuelle de la lésion. Le dermatologue, à l'aide d'un dermatoscope, peut analyser la peau et détecter les différents indicateurs de cancer. Parmi les différentes manières de détecter les cancers, il existe plusieurs règles contenant différents critères pour détecter les lésions dangereuses. La règle la plus utilisée est celle dite ABCDE, que nous allons décrire ci-dessous, et qui signifie : Asymétrie, Bords irréguliers, Couleurs

différentes, Diamètre et Évolution. Si une lésion représente un ou plusieurs de ces critères, alors la possibilité qu'il s'agisse d'un cancer augmente considérablement. Un autre concept tout aussi important, mais encore peu étudié dans la recherche, est le concept du 'Vilain Petit Canard', que nous définirons également dans les parties suivantes. Toutefois, le diagnostic n'est pas aisé. Selon plusieurs études, les dermatologues parviennent à diagnostiquer correctement seulement 74 % des lésions cutanées des patients lors d'analyses visuelles [11]. Nous verrons dans les parties suivantes comment l'intelligence artificielle peut permettre d'améliorer le diagnostic à l'aide de méthodes automatisées d'assistance au diagnostic.

1.1.7.1. Règle "ABCDE"

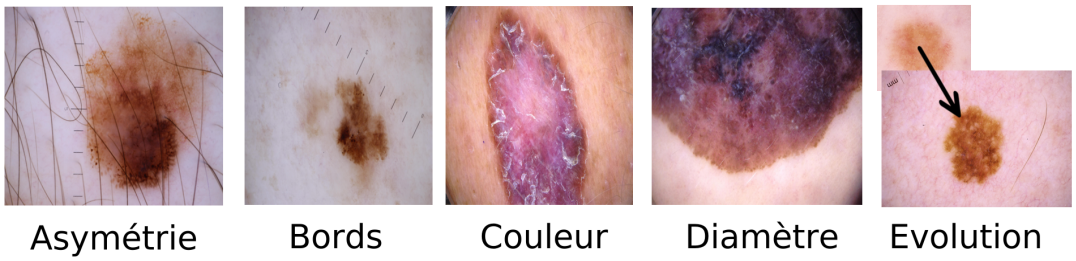

FIGURE 1.7. – **Schéma représentant les différents critères de la règle ABCDE.**

La règle ABCDE [12] occupe une place prépondérante dans les méthodes d'aide au diagnostic du mélanome. Cette règle, qui se décompose en cinq critères principaux, permet aux professionnels de la santé et au grand public de surveiller les lésions cutanées et de détecter les changements qui pourraient indiquer la présence d'un mélanome. La Figure 1.7 présente les différents critères, illustrés par des mélanomes présentant les caractéristiques correspondantes.

Asymétrie Le premier critère de la règle ABCDE est l'asymétrie, qui fait référence à la façon dont les moitiés de la lésion ne correspondent pas symétriquement lorsqu'on les divise en deux. Les mélanomes ont tendance à être asymétriques, contrairement aux lésions bénignes qui ont souvent une bonne symétrie. Cette asymétrie peut être observée visuellement en comparant les deux côtés de la lésion ou en traçant une ligne médiane à travers celle-ci.

Bordure Le deuxième critère, la bordure, fait référence à la façon dont les bords de la lésion sont définis. Les mélanomes ont souvent des bords irréguliers, dentelés ou mal définis, tandis que les lésions bénignes ont tendance à avoir des bords lisses et réguliers. Cette irrégularité dans la bordure peut être un signe de malignité et devrait inciter à une évaluation plus approfondie par un professionnel de la santé.

Couleur La couleur de la lésion est un autre critère important de la règle ABCDE. Les mélanomes peuvent présenter une variété de couleurs, allant du brun foncé au noir, en passant par des nuances de rouge, de bleu ou de blanc. La présence de multiples couleurs ou de variations de teintes dans une lésion peut être un indicateur de mélanome et devrait être surveillée de près.

Diamètre Le quatrième critère de la règle ABCDE est le diamètre de la lésion. Bien que les mélanomes aient généralement un diamètre supérieur à 6 millimètres, il est important de noter que des lésions plus petites peuvent également être malignes, en particulier si elles présentent d'autres caractéristiques suspectes. Il est donc essentiel de surveiller toutes les lésions cutanées, quel que soit leur diamètre.

Evolution Enfin, le dernier critère de la règle ABCDE est l'évolution de la lésion au fil du temps. Les changements dans la taille, la forme, la couleur ou la texture d'une lésion cutanée existante peuvent être un signe de mélanome. Il est crucial de surveiller toute évolution dans une lésion cutanée et de consulter un professionnel de la santé si des changements sont observés.

1.1.7.2. Le concept du “Vilain Petit Canard” (VPC)

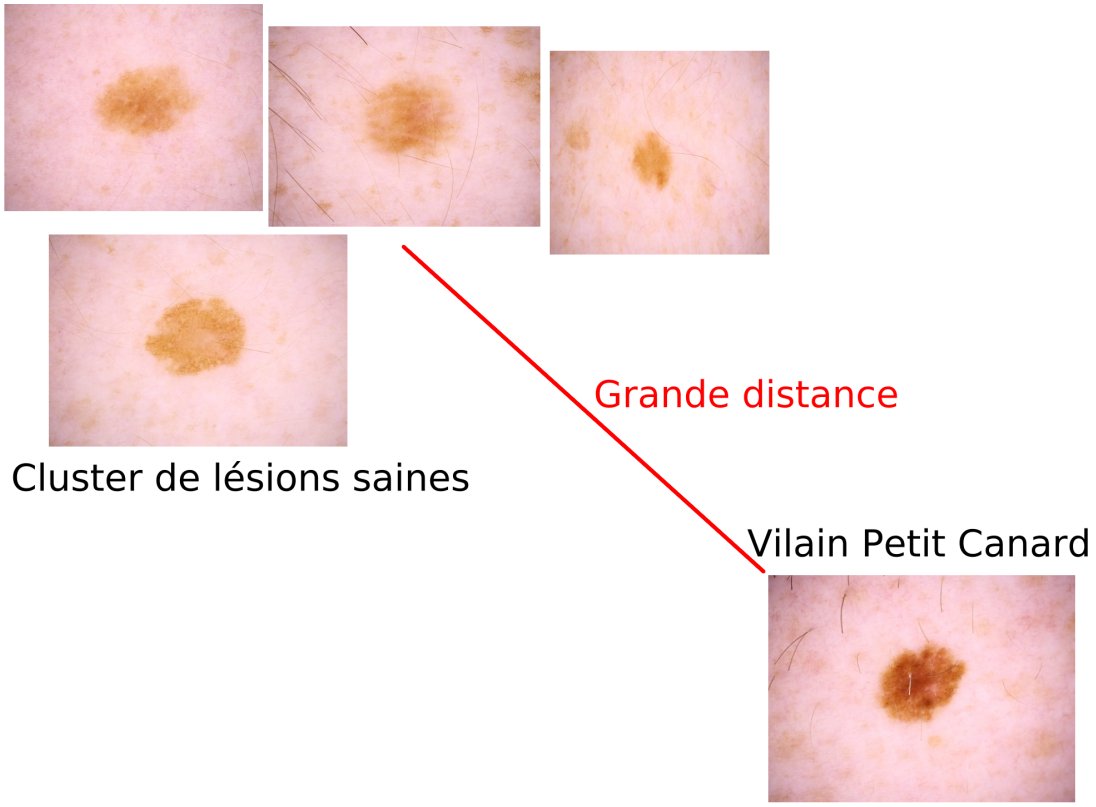

FIGURE 1.8. – **Schéma représentant le concept du Vilain Petit Canard pour le diagnostic du mélanome.** Intuitivement, cela correspond à la détection d’une lésions très différente des autres. Informatiquement, cela se traduira par une grande distance dans l’espace de représentations.

Le concept du vilain petit canard, pour la première fois proposé par Jean-Jacques Grob en 1998 [13], repose sur l’idée que, pour un patient donné, les mélanomes se distinguent souvent comme les "vilains petits canards" en raison de leur apparence atypique par rapport aux autres lésions présentes chez le patient. Ce concept s’ajoute aux autres règles existantes en apportant une approche différente : la comparaison avec les autres lésions cutanées du patient.

Cette approche originale permet non seulement un meilleur diagnostic pour les dermatologues, mais offre également une perspective nouvelle pour les méthodes de diagnostic assistées par ordinateur, qui, historiquement, traitaient les images de manière isolée. En intégrant l’information relative aux autres lésions du patient, il est probable que les performances des méthodes existantes pour l’aide au diagnostic soient améliorées, tout en permettant le développement de méthodes d’interprétabilité plus avancées.

Cependant, nous verrons par la suite que ces approches sont encore récentes et que

la meilleure manière d'utiliser ce concept n'est pas encore clairement définie.

1.1.7.3. Les informations patient

Enfin, le dernier outil que nous présentons est l'utilisation des informations relatives au patient. Celles-ci incluent des données classiques telles que l'âge, le sexe, mais aussi des informations plus détaillées comme les antécédents de coups de soleil, les habitudes d'exposition au soleil du patient, ainsi que son utilisation de protection solaire ou de bancs UV. De manière générale, il est recommandé que le dermatologue adopte une approche holistique, en considérant le patient dans sa globalité pour établir un diagnostic. Cette démarche est également liée à plusieurs facteurs statistiques connus en dermatologie, certaines populations étant plus susceptibles de développer un mélanome.

1.2. Cadre Méthodologique

Ces méthodes ont rapidement été appliquées à divers problèmes médicaux, y compris le diagnostic du mélanome. Pour ce faire, il a fallu créer des bases de données, définir des métriques, et établir un cadre méthodologique complet que nous allons décrire dans cette section. Tout d'abord, nous présenterons les bases de données qui ont émergé au fil des années, suivies des diverses métriques utilisées pour mesurer les performances des modèles. Ensuite, nous aborderons les premiers travaux réalisés en matière de classification du mélanome et de détection d'anomalies.

1.2.1. Bases de données dermatologiques

Concernant les bases de données, ces dernières se sont avant tout focalisées sur l'usage d'images dermoscopiques des lésions cutanées. Ces photographies, prises avec un dermatoscope directement appliqué sur la peau du patient, offrent une vue détaillée et de haute résolution de la lésion examinée.

Les différentes bases de données Il existe plusieurs bases de données de lésions cutanées dont les principales sont présentées Table 1.1. Les premières bases de données ne contenaient que quelques centaines d'images, ce qui rendait l'apprentissage automatique presque impossible. En 2013, la base de données Dermofit Image Library [14] a atteint le millier d'images, puis, à partir de 2016, les bases de données de l'ISIC se sont imposées comme la norme en termes de bases de données pour la détection du mélanome, grâce à leur nombre conséquent d'images. La première version, en 2016 [15], contenait 900 images, puis ce nombre n'a cessé d'augmenter, passant à 2 000 pour ISIC 2017 [16], 10 015 pour ISIC 2018 [17, 18], 25 331 pour ISIC 2019 [19-21], et enfin 33 116 avec ISIC 2020 [3]. Toutes ces bases sont annotées, mais il est important de noter que les annotations n'ont pas toujours été les mêmes. Dans

le cas d'ISIC 2019, par exemple, les annotations classifiaient les images en 8 catégories (Mélanome, Nævus mélanocytaire, Carcinome basocellulaire, Kératose actinique, Kératose bénigne (lentigo solaire / kératose séborrhéique / kératose lichénoïde), Dermatofibrome, Lésion vasculaire, Carcinome épidermoïde, et la catégorie Aucun des éléments ci-dessus), tandis que d'autres bases de données regroupaient les images en mélanome et nævus. Un autre point important concerne la méthode d'annotation des images : alors que tous les mélanomes ont été analysés par biopsie, garantissant ainsi la certitude de leur malignité, les nævus ont simplement été annotés selon l'évaluation des dermatologues, qui n'ont pas jugé nécessaire de réaliser une biopsie. Bien que cette approche soit tout à fait légitime afin de ne pas perturber les patients, il est possible que de nombreux mélanomes aient été annotés comme bénins dans les bases de données ISIC.

Par ailleurs, ces bases de données regroupent en réalité différentes collections provenant de plusieurs hôpitaux à travers le monde. Les deux dernières versions contiennent également de nombreuses métadonnées, qui sont très utiles pour les réseaux de neurones. Ces informations patients correspondent à l'âge, le sexe du patient, la localisation de la lésion sur le corps, le diagnostic réel, et contiennent également un identifiant patient permettant de regrouper les images par patient, utile pour la détection d'anomalies et du traitement du concept du vilain petit canard.

Nom	Mélanomes	Total	Méta-données
MClass-D Dataset [1]	80	100	×
PH2 Database [2]	40	200	×
Dermofit Image Library [3]	76	1 300	×
ISIC 2016 [4]	173	900	×
ISIC 2017 [5]	374	2 000	×
ISIC 2018 [6]	1 113	10 015	×
ISIC 2019 [7]	4 522	25 331	✓
SIIM-ISIC 2020 [8]	584	33 126	✓

TABLEAU 1.1. – **Présentation des principales bases de données d'images dermoscopiques, affichées par nombre croissant d'images.** [1][22] [2][23] [3][14] [4][15] [5][16] [6][17, 18] [7][19-21] [8][3]

En particulier, la base SIIM-ISIC 2020 se démarque par sa quantité de données bien plus grande que les autres, qui est un critère fondamental en apprentissage automatique. Naturellement, nos différents travaux se sont basés sur cette base de données publique, disponible librement sur internet. La majorité des travaux présents dans la littérature utilisent également cette base de données. Toutefois, en fonction du contexte, certaines approches utilisent également les autres versions d'ISIC pour leurs spécificités. ISIC 2018 contient de nombreuses classes en plus du traditionnel mélanome et naevus, et ISIC 2019 contient substantiellement plus d'images de mélanomes. Par exemple, dans le cas de la compétition en ligne d'ISIC 2020, nombreux étaient les participants à entraîner leurs modèles sur l'ensemble des bases existantes, et non pas

sur une seule base.

Présentation d'ISIC 2020 Ainsi, la base de données que nous avons utilisé tout au long de nos expériences est ISIC-2020. Cette dernière, en plus d'être la plus grande en termes de nombre d'images, contient également des informations essentielles pour l'utilisation du concept du VPC : l'identifiant du patient correspondant à chaque lésion. Avec cette information, nous pouvons regrouper les images par patient, les comparer et les analyser afin de détecter quelles lésions sont atypiques par rapport aux autres.

Comme toutes les autres bases de données dermatologiques, ISIC-2020 est très déséquilibrée, elle contient environ 98% de naevus pour seulement 2% de mélanomes. Pour chacun de nos travaux, nous détaillerons les mesures que nous avons prises en considération afin que ce déséquilibre ne perturbe pas nos méthodes.

Concrètement, cette base regroupe plus de 32 000 images, toutes annotées entre mélanome et naevus. Ce sont des images dermoscopiques prises à l'aide d'un dermatoscope photoluminescent, qui, bien que la majorité soient de bonne qualité, de nombreuses images présentent tout de même des artefacts tels des poils devant la lésion, des bulles d'airs visibles, de l'encre, et tout autre éléments perturbateurs cachant une partie des lésions. Ces artefacts sont présents dans certaines images, et peuvent négativement impacter les performances des modèles. Nous verrons les méthodes d'augmentations de données que nous avons utilisé pour rendre nos modèles plus robustes aux artefacts.

La création de cette base de données par plusieurs hôpitaux durant l'exercice clinique est très utile car elle contient un très grand nombre d'images... mais cette diversité apporte son lot de défis, notamment en ce qui concerne la colorimétrie des images. Certaines présentent une teinte bleutée, tandis que d'autres sont plus rougeâtres. Ces variations de couleur n'affectent pas l'analyse effectuée par un dermatologue, mais elles peuvent influencer les modèles d'intelligence artificielle qui, sans information sur l'origine des images, tentent de déterminer si la couleur influence le diagnostic. Heureusement, des études ont montré que cet impact peut être négligeable, surtout avec l'application de techniques appropriées de data augmentation.

Vers l'usage de reconstructions 3D Enfin, de nouvelles méthodes d'imagerie tendent à se démocratiser afin de rendre la capture de données encore plus rapide et efficace. Par exemple, l'utilisation des machines de reconstruction 3D s'accroît rapidement dans les services dermatologiques à travers le monde. Ces machines permettent la prise d'un très grand nombre de photos simultanément, tout autour du patient et sous différents angles, afin de reconstruire un modèle en trois dimensions de sa peau. Une fois le modèle 3D généré, des méthodes de prétraitement sont ensuite utilisées pour détecter les lésions, les isoler puis les soumettre à des algorithmes d'aide au diagnostic. Peu de services dermatologiques ont déjà accès à ces machines, mais leur usage devient de plus en plus fréquent, et il est fort probable que de nombreux travaux de recherche se focaliseront sur cette nouvelle modalité afin de développer

des outils pour les dermatologues.

En France, certains services dermatologiques possèdent déjà ce genre de machine, permettant la génération de modèles 3D des patient avec une bonne résolution de la peau des patients.

1.2.2. Métriques utilisées

Afin de mesurer précisément les performances des modèles, il est nécessaire de définir plusieurs types de métriques adaptées à notre problème, que nous allons décrire juste ici.

Accuracy Dans le cadre de la classification, une approche classique consiste à analyser l'accuracy (aussi appelée précision en français). Cette dernière est simplement calculée en divisant le nombre d'images correctement classées $n_{correct}$ par le nombre total d'images n_{total}, selon la formule suivante :

$$Accuracy = \frac{n_{correct}}{n_{total}}$$

Cette métrique présente toutefois un problème majeur lorsque nous travaillons avec des bases de données déséquilibrées. Si le nombre d'images saines est significativement supérieur au nombre d'images de cancers, alors une accuracy apparemment élevée peut faussement suggérer de bonnes performances. Un nombre faible d'images de cancers influencera peu l'accuracy globale, par rapport aux nombreuses images de lésions saines. Il est donc nécessaire d'utiliser d'autres métriques pour évaluer correctement les performances du modèle, de sorte à ce que les instances de cancers aient une importance élevées lors du calcul des performances.

Courbe ROC Afin d'obtenir plus d'informations sur les prédictions générées par les modèles, la courbe ROC (Receiver Operating Characteristic en anglais) permet de représenter la performance du modèle à différents seuils de classification. Cette courbe illustre le taux de vrais positifs (sensibilité) par rapport au taux de faux positifs (1 - spécificité), offrant ainsi une évaluation visuelle de la capacité du modèle à distinguer les classes de manière efficace. L'aire sous la courbe ROC (AUC) peut également être calculée pour quantifier la performance globale du modèle, indépendamment des seuils choisis. Un modèle complètement aléatoire aurait un score AUC de 0,5, tandis qu'un modèle parfait atteindrait un score de 1.

Sensibilité et spécificité Une fois le seuil idéal déterminé grâce à la courbe ROC, il est possible de calculer la sensibilité et la spécificité correspondantes. La sensibilité mesure la proportion de vrais positifs détectés parmi les cas effectivement positifs, tandis que la spécificité évalue la capacité du modèle à identifier correctement les cas négatifs. Dans le domaine médical, par exemple, il est souvent crucial d'obtenir de bonnes performances sur les cas positifs, qui représentent les maladies.

Il convient toutefois de noter que, bien que ces deux métriques soient plus stables et plus représentatives de la véritable performance du modèle, elles ne sont pas totalement à l'abri des problèmes liés au déséquilibre des classes. En effet, une AUC élevée peut aussi être biaisée par une grande proportion d'images saines.

Score F1 Le score F1 est une métrique cruciale, souvent utilisée en complément ou en alternative à la sensibilité et à la spécificité, en particulier dans les contextes où les classes sont déséquilibrées. Ce score représente la moyenne harmonique de la précision et du rappel, ce qui permet d'évaluer l'équilibre entre ces deux métriques. Une valeur élevée du score F1 indique non seulement que le modèle détecte efficacement les cas positifs (rappel élevé), mais qu'il le fait également avec une grande précision, minimisant ainsi le nombre de faux positifs parmi les cas positifs détectés. La formule pour calculer le score F1 est la suivante :

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

où :

— **Precision** est le ratio des vrais positifs par rapport à tous les positifs prédits (vrais positifs + faux positifs),

— **Recall** est le ratio des vrais positifs par rapport à tous les cas réellement positifs (vrais positifs + faux négatifs).

Dans des domaines comme la détection de fraudes ou le diagnostic médical, où il est crucial de minimiser à la fois les faux négatifs et les faux positifs, le score F1 s'avère particulièrement utile. Il fournit une vue plus complète de la performance du modèle que les mesures isolées de précision ou de rappel, rendant ainsi la métrique particulièrement adaptée pour les applications où les erreurs ont des conséquences significatives. Le score F1 aide également à identifier les cas de performances apparemment élevées qui pourraient en réalité être trompeuses en raison du déséquilibre des classes, offrant ainsi une mesure plus fiable de l'efficacité réelle du modèle.

Courbe précision-rappel Enfin, la courbe précision-rappel, bien que moins fréquemment utilisée, est un outil essentiel pour évaluer la performance des modèles de classification dans des contextes où les classes sont fortement déséquilibrées. Elle illustre le compromis entre la précision, qui mesure la proportion de vrais positifs parmi les prédictions positives, et le rappel, également connu sous le nom de sensibilité, qui quantifie la proportion de vrais positifs détectés parmi les cas réellement positifs. D'une manière assez similaire à la courbe ROC, elle est particulièrement utile pour comprendre comment la performance du modèle varie avec différents seuils de classification, permettant ainsi aux chercheurs de choisir un seuil qui optimise l'équilibre entre précision et rappel selon les besoins spécifiques de leur application. L'aire sous la courbe de la courbe précision-rappel peut également servir à évaluer la qualité globale du modèle, avec des valeurs plus élevées indiquant une meilleure performance générale.

Toutefois, cette métrique n'est encore que très peu utilisée dans la littérature, et il est rare de voir des articles incluant cette métrique. Ainsi, la majorité des articles sur le diagnostic du mélanome se concentrent sur l'AUC, la sensibilité et la spécificité.

1.3. Travaux précédents sur la classification de mélanomes

Les connaissances dermatologiques décrites dans la partie précédente ont longtemps permis aux dermatologues d'améliorer leur diagnostic. Depuis l'usage popularisé de l'informatique, ces approches n'ont cessé d'être automatisées et adaptées à des logiciels et algorithmes de diagnostic. Dans cette partie, nous allons présenter les différentes approches qui ont été utilisées pour détecter les mélanomes, et de manière plus générale, nous introduisons ici le cadre général de la vision par ordinateur, des premières approches à base de systèmes experts aux méthodes récentes à base de réseaux de neurones.

1.3.1. Introduction

La médecine a toujours été alimentée par les nouvelles avancées technologiques des autres domaines tels que les mathématiques, la physique ou encore l'informatique. Le développement de ces sciences, appliquées à la médecine, a permis de nombreuses innovations qui aujourd'hui permettent de détecter des maladies, créer des médicaments et soigner au mieux les malades. Les méthodes d'imageries médicales proviennent pour la majorité d'innovations de la physique, ainsi que des mathématiques. De manière un peu plus récente, l'informatique a également permis son lot de contributions à la médecine, à l'aide d'algorithmes d'intelligences artificielles qui permettent de faire gagner toujours plus de temps aux médecins, mais également de rendre les diagnostics et prises en charge plus rapides et précis. L'impact clinique est réel, et son impact sur la recherche en médecine grandit également petit à petit. Dans cette partie, nous étudions en détail comment ces innovations sont implémentées pour le diagnostic du mélanome. Nous discutons des travaux passés, mais également des travaux présents, et de ce vers quoi tend la recherche actuelle dans ce domaine.

1.3.2. Les systèmes experts en vision par ordinateur

Bien avant les méthodes actuelles d'apprentissage automatique, le diagnostic assisté par ordinateur reposait sur des ensembles de critères définis par des experts pour détecter les maladies. Dans le cas du mélanome, ces approches étaient basées sur des règles telles que la règle "ABCDE" décrite précédemment [24, 25]. Ces méthodes étaient, de manière générale, très complexes à mettre en œuvre pour plusieurs raisons. Tout d'abord, il était nécessaire que les dermatologues et les informaticiens collaborent étroitement afin de définir des critères réellement importants pour le diagnostic. Ensuite, l'implémentation de ces critères n'était pas toujours évidente : un

concept simple pour l'humain peut être extrêmement compliqué à mettre en place informatiquement. En ce qui concerne les images, ce problème est d'autant plus crucial, car la manière dont les images sont présentées dans un ordinateur ne permet pas toujours d'effectuer des calculs simples et efficaces pour en extraire des caractéristiques. Détecter une forme ou un patron sur une lésion requiert la détection de pixels suivant une certaine logique, l'analyse des pixels environnants et la prise en compte de tous les cas de figure. Ces approches sont donc très sensibles au bruit et à toute sorte de variations imprévues dans les images. Toutefois, leurs avantages résident dans le fait qu'elles consomment généralement moins d'énergie que les modèles profonds, et que leurs prédictions sont plus facilement compréhensibles et interprétables que celles des réseaux de neurones.

En terme de modèles utilisés pour ces approches, il s'agit principalement de Machines à Vecteurs de Support (SVM) [26, 27], d'Arbres décisionnels et de Forêts Aléatoires [28, 29].

1.3.3. Réseaux de neurones artificiels

Malgré la puissance des approches précédentes, celles basées sur les réseaux de neurones profonds se sont démarquées par leur efficacité et leurs succès dans de nombreux domaines jusque-là trop difficiles pour les algorithmes classiques. Les réseaux de neurones, grâce à leur capacité à générer des caractéristiques complexes à partir de simples annotations, ont permis non seulement d'améliorer grandement les performances dans les tâches de classification d'images, mais aussi de résoudre une vaste diversité de problèmes tels que la segmentation, la reconstruction et la génération de nouvelles instances.

Concernant leur mode de fonctionnement, les réseaux de neurones sont des algorithmes capables d'accomplir des tâches spécifiques en se basant sur un processus d'apprentissage supervisé. Au cours de la phase d'entraînement, le modèle ajuste progressivement les poids de ses neurones afin de produire une sortie qui correspond à l'objectif souhaité. Une fois l'entraînement terminé, le modèle est théoriquement capable de fournir des prédictions précises en fonction des données en entrée. Dans le contexte du diagnostic du mélanome, ces prédictions prennent la forme d'une classification binaire, indiquant soit la présence d'un mélanome, soit celle d'une lésion bénigne.

Contrairement aux approches traditionnelles qui nécessitent la définition explicite de critères ou de caractéristiques spécifiques, les réseaux de neurones apprennent automatiquement à extraire et à utiliser les caractéristiques pertinentes à partir des données brutes, réduisant ainsi la dépendance aux calculs préalables manuels. Un schéma est présenté Figure 1.9.

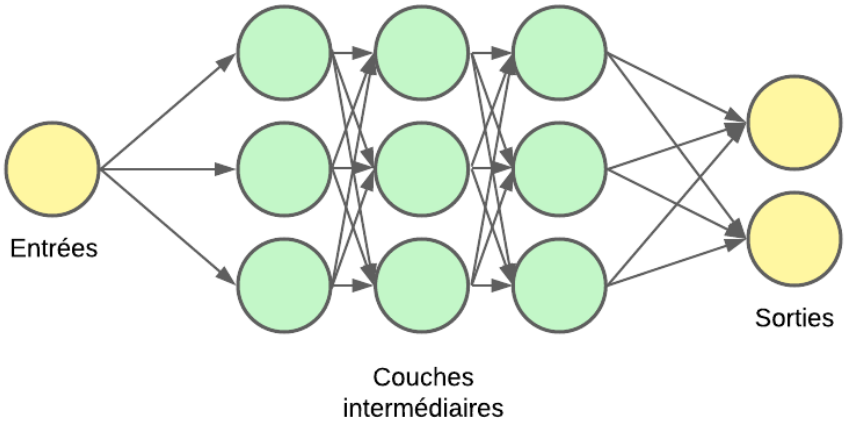

FIGURE 1.9. – **Schéma simplifié d'un réseau de neurones à quatre couches.**

Au fil des années, une multitude de réseaux de neurones différents ont émergé, chacun doté de particularités le rendant plus performant pour des tâches spécifiques. Les réseaux de neurones convolutifs (CNN) sont efficaces pour la classification d'images, tandis que les Generative Adversarial Networks (GAN) sont utilisés pour la génération d'images et les réseaux siamois pour la comparaison d'images, entre autres. Bien que tous ces réseaux soient des variantes de neurones, ils présentent des différences significatives dans leur architecture ou leurs composants.

En ce qui concerne la classification d'images, ces avancées ont permis une amélioration fulgurante des performances ces dernières années. Par exemple, lors des compétitions de classification d'images telles qu'ImageNet (qui est l'une des plus grandes base de données annotées d'images), les meilleurs modèles sont tous des réseaux de neurones depuis 2013 (Voir figure 1.10).

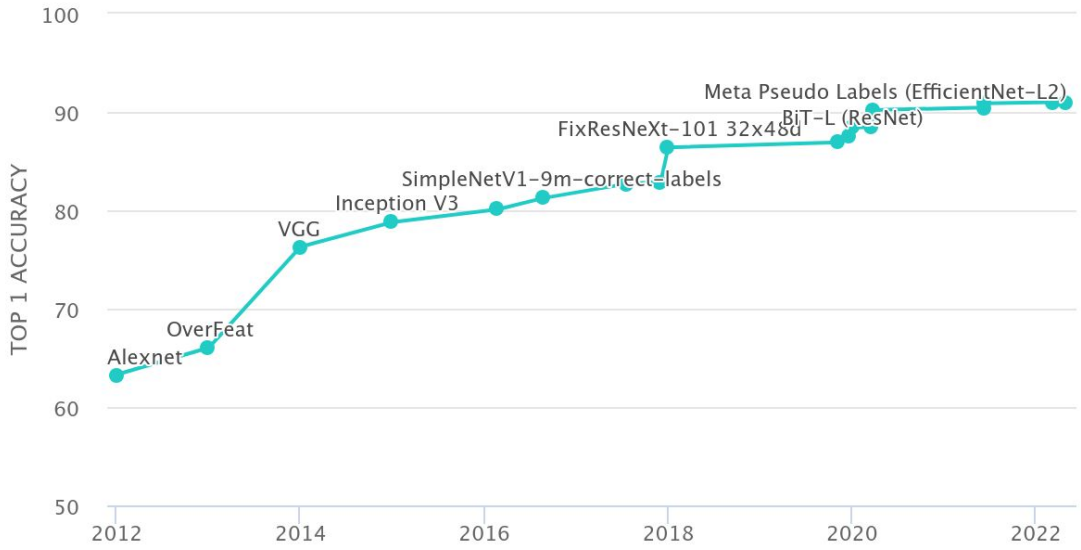

FIGURE 1.10. – **Évolution des performances de classification d'images au fil des années sur la base de données ImageNet [30].** En 2013, le réseau de neurones convolutif AlexNet [31] a grandement amélioré les performances. *Crédit image : [32]*

De plus, de nombreux composants des réseaux de neurones sont personnalisables comme les fonctions d'activation, les optimisateurs, le nombre de neurones... toutes ces caractéristiques font que deux CNN peuvent porter le même nom mais avoir d'innombrables caractéristiques différentes l'un de l'autre. Parmi cette diversité de réseaux, nous nous concentrerons ici sur les CNN et les réseaux siamois qui sont les types de modèles qui nous intéressent pour la classification et la comparaison d'images.

L'entrée du réseau de neurones est ce que l'on va lui donner pour qu'il puisse réaliser sa tâche. Celle-ci peut être une image, que l'on désire classer parmi les catégories "chat" et "chien", ou encore un texte que l'on veut classer dans les styles "roman" et "essai". De manière générale, on peut imaginer n'importe quel type de données, comme des vidéos, de l'audio, des coordonnées GPS, etc.

Un réseau de neurones est composé d'une ou plusieurs couches, chacune composée d'un certain nombre de neurones. Concrètement, les neurones sont des valeurs répertoriées sous forme de matrices que l'on appelle "poids". D'abord aléatoires, ces valeurs vont, au fil de l'entraînement, reproduire les résultats demandés à l'aide de deux phases : la propagation avant (forward pass), qui calcule les sorties du modèle, et la rétropropagation du gradient (backpropagation), qui corrige ces sorties et met à jour les poids du modèle.

1.3.4. Applications au mélanome

Les premières approches d'intelligence artificielle pour le diagnostic du mélanome remontent au début des années 1990, avec des travaux visant tout d'abord à segmenter

les lésions et à analyser leurs couleurs [33-35]. Ces approches reposaient entièrement sur des systèmes experts, dans lesquels les caractéristiques étaient sélectionnées manuellement pour détecter les mélanomes. Pendant deux décennies, ces approches ont dominé le domaine, produisant des résultats convaincants pour l'époque. Une multitude de caractéristiques inspirées par les critères dermatologiques ont ainsi été développées, chacune utilisée pour entraîner les systèmes experts. Malgré tout, le problème principal de ces méthodes est qu'elles sont très sensibles aux bruits, aux variations de couleurs et de luminosité ainsi qu'aux artefacts (éléments perturbant l'analyse des lésions, tels que les poils ou le bruit des appareils d'imagerie). Par conséquent, leur utilisation en clinique restait limitée et marginale.

Par la suite, les réseaux de neurones ont offert une approche plus robuste pour le diagnostic du mélanome grâce à l'entraînement de réseaux convolutifs (CNN) sur des bases de données étiquetées. Les premières approches ont été développées en s'appuyant sur diverses architectures préconstruites, telles que ResNet [36], VGG [37] ou EfficientNet [38], ainsi qu'en créant de nouvelles architectures [39], produisant des résultats très encourageants par rapport aux méthodes antérieures. Cette grande diversité de travaux a été analysée dans plusieurs études comparatives [40, 41], qui concluent de manière unanime que les modèles profonds sont capables de détecter les mélanomes avec des performances impressionnantes, atteignant des AUC entre 0,80 et 0,94, là où les systèmes experts peinaient à dépasser 0,80. Cependant, malgré ces bonnes performances, l'intégration de ces méthodes en clinique peine encore à s'imposer dans les examens de routine, en partie en raison du manque d'explicabilité des décisions prises par ces modèles.

Plus récemment, et afin de rendre les résultats plus interprétables, de plus en plus d'études tentent de combiner les systèmes experts et les réseaux de neurones, comme dans les travaux de [42], [43], [44] ou encore [45]. D'une part, l'approche "système expert" se base sur les connaissances dermatologiques afin de rendre les résultats interprétable et d'autre part, les approches profondes permettent une meilleure robustesse et de meilleures performances. Leur fusion permet ainsi un potentiel gain en performance et en interprétabilité.

Il existe également des méthodes entièrement à base de réseaux de neurones mais dont l'architecture est augmentée par les connaissances dermatologiques, par exemple en créant un graphe composés de réseaux de neurones où chaque réseau est spécialisé dans une maladie spécifique [46].

Concernant l'usage en clinique, des études pionnières, telles que celles de [47] et [48], ont démontré que les CNN pouvaient atteindre une précision comparable, voire supérieure, à celle des dermatologues expérimentés, en utilisant de grandes bases de données d'images dermoscopiques. D'autres recherches importantes incluent l'étude de [49], qui a utilisé des ensembles de données diversifiés pour améliorer la généralisation des modèles, et [50], qui a exploré la collaboration entre dermatologistes et intelligence artificielles afin d'améliorer le diagnostic.

Ces avancées ont considérablement amélioré la précision du diagnostic du mélanome, en exploitant des volumes importants de données et en développant des architectures de réseau toujours plus sophistiquées. Cependant, ces approches se

heurtent également à plusieurs défis, dont deux particulièrement importants :

— **L'usage d'images uniques.** La quasi-totalité des articles présents dans la littérature effectuent leurs prédictions sur une image à la fois, ce qui pose problème en raison de la nature du mélanome et de sa ressemblance avec les nævus. Comme nous l'avons décrit précédemment, la comparaison des lésions entre elles est une étape importante du diagnostic, notamment à travers le concept du vilain petit canard. Les modèles classiques fonctionnent tous exclusivement en prenant une seule image en entrée, ce qui peut être insuffisant pour réaliser un diagnostic efficace. Toutefois, encore aujourd'hui, peu de travaux dans la littérature scientifique abordent réellement la comparaison des lésions intra-patient tout en obtenant de bons résultats. Dans la section suivante, nous présenterons en détail les articles récents traitant de cette problématique ainsi que les défis associés.

— **La difficulté à comprendre les prédictions de ces modèles.** Le second problème majeur des approches actuelles est le manque d'interprétabilité, qui soulève également une question éthique cruciale quant à la prise de décision finale dans un diagnostic. Dans le milieu médical en particulier, comprendre les décisions relatives au diagnostic est d'une importance capitale afin d'éviter les erreurs médicales. En confiant la responsabilité des décisions à des algorithmes sans les comprendre pleinement, nous risquons de commettre des erreurs de diagnostic évidentes pour les dermatologues mais que la machine n'a pas détecté. L'opacité des modèles de deep learning, souvent qualifiés de "boîtes noires", limite la confiance des médecins et des patients dans les décisions automatisées, et il est donc essentiel de développer des méthodes d'explicabilité et de transparence pour ces modèles.

1.3.5. Approches par comparaison d'images

La comparaison d'images est intrinsèquement plus complexe que la simple classification d'images. Alors que la classification peut s'appuyer sur des annotations basiques décrivant le contenu de l'image, la comparaison nécessite une compréhension approfondie des caractéristiques visuelles pour comparer les images sur plusieurs critères. Cette complexité se reflète à la fois dans l'annotation et dans la conception des modèles.

Dans le cadre de la comparaison d'images, les lésions doivent être évaluées selon des critères tels que l'asymétrie ou la régularité des bords, qui sont souvent subjectifs et varient d'un dermatologue à l'autre. Proposer une échelle d'annotation nuancée (par exemple, une évaluation des critères sur une échelle de 1 à 100) s'avère complexe et chronophage. Actuellement, il existe un manque de bases de données annotées avec des concepts spécifiques, ce qui oblige de nombreux modèles à opérer sans ces annotations, en se limitant à des étiquettes simples comme "mélanome" ou "naevus".

Cette absence d'annotations basées sur des concepts freine le développement de nouvelles méthodes pourtant très efficaces dans d'autres domaines. Par exemple, de nombreux articles en apprentissage automatique pour le biomédical utilisent les

rapports écrits des cliniciens associés aux images médicales pour entraîner des architectures spécifiques reliant le texte et l'image. C'est notamment le cas de l'architecture CLIP [51], qui place les images dans un espace latent multimodal, créant des liens entre textes et images. S'inspirer de CLIP pourrait nous conduire à développer une base de données avec des annotations textuelles telles que "asymétrique", "bords irréguliers" ou "différences de couleur". Cela permettrait non seulement d'exploiter la puissance de CLIP pour le diagnostic, mais aussi de représenter les images de lésions dans un espace latent facilement interprétable, améliorant ainsi notre compréhension des modèles. En raison de cette limite, nous nous sommes orientés vers d'autres approches, mais il est évident que ces nouvelles méthodes seront un jour intégrées dans le cadre de la détection du mélanome.

Enfin, il existe également d'autres méthodes innovantes, encore peu exploitées dans le contexte du mélanome, telles que la comparaison d'images réalisée sans annotations, en utilisant diverses méthodes auto-supervisées sur les bases de données existantes. Dans les chapitres suivants, nous étudierons en détail ces approches actuelles, ainsi que les méthodes employées pour contourner les défis que nous avons soulevés.

1.4. Synthèse

Les mélanocytes sont des cellules responsables de la production de mélanine, le pigment qui donne la couleur à notre peau. Cependant, certains facteurs endogènes, comme des prédispositions génétiques, ou exogènes, tels que l'exposition excessive aux rayons UV, augmentent les risques de transformation de ces cellules en mélanome, un cancer de la peau particulièrement agressif. Le diagnostic du mélanome est complexe en raison de sa ressemblance avec des lésions cutanées bénignes.

Pour faciliter le diagnostic, plusieurs méthodes cliniques sont utilisées, comme la règle ABCDE qui se base sur des critères de la lésion, ainsi que le concept du "vilain petit canard" qui consiste à identifier une lésion qui se distingue des autres grains de beauté du patient. Malgré ces outils, le diagnostic demeure un défi et peut bénéficier des avancées technologiques.

Les méthodes d'apprentissage profond, telles que les CNN, offrent des perspectives prometteuses pour l'automatisation de la détection des mélanomes. Ces technologies analysent les images de lésions cutanées pour détecter des caractéristiques permettant de classer les images entre mélanome et grains de beauté. En intégrant ces méthodes automatisées dans les pratiques cliniques, il est possible d'améliorer la précision du diagnostic et de détecter les mélanomes à un stade plus précoce, ce qui est crucial pour le pronostic et le traitement des patients.

Ces dernières décennies, les améliorations apportées aux modèles profonds ont permis d'atteindre des performances remarquables sur une grande diversité de bases de données pour la classification d'images. Dans le domaine du mélanome, cette évolution s'est également manifestée par des résultats toujours meilleurs. Toutefois, nous remarquons que les modèles actuels présentent d'importantes faiblesses qui

peuvent nuire à leurs performances. L'usage simultané de plusieurs images est une approche très intéressante pour la détection du mélanome, mais elle reste encore trop peu étudiée. Les approches récentes basées sur l'attention et les Transformers, ainsi que de nouvelles méthodes, pourraient pallier ces problèmes et permettre d'améliorer encore les performances ainsi que l'interprétabilité des modèles.

Nous explorerons ces méthodes automatisées visant à détecter les mélanomes dans la section suivante. Ces approches pourraient révolutionner la manière dont les mélanomes sont diagnostiqués, offrant ainsi de meilleures chances de survie aux patients grâce à une détection précoce et précise.

2. Comparaison des lésions intra-patient pour la détection du mélanome

Sommaire

2.1. Introduction . . . 50
2.2. Revue de la littérature . . . 50
2.2.1. Caractérisation du VPC par des participants . . . 51
2.2.2. Analyse de la perte d'AutoEncodeurs . . . 51
2.2.3. Calcul des distances dans l'espace latent . . . 53
2.2.4. Approches basées sur l'attention . . . 54
2.2.5. Méthodes de détection d'anomalies . . . 55
2.3. Détection non supervisée des mélanomes . . . 55
2.3.1. Introduction . . . 55
2.3.2. Méthode . . . 56
2.3.2.1. Architecture globale . . . 56
2.3.2.2. L'architecture SimSiam comme encodeur . . . 57
2.3.2.3. Une variante des k-plus-proches voisins pour comparer les images . . . 58
2.3.3. Résultats . . . 60
2.3.4. Conclusion . . . 62
2.4. Détection supervisée du mélanome par comparaison de lésions cutanées par patient . . . 63
2.4.1. Introduction . . . 63
2.4.2. Méthode . . . 64
2.4.2.1. ReSeT . . . 64
2.4.2.2. Encodeur . . . 65
2.4.2.3. Comparaison d'images à base de SetTransformers . . . 66
2.4.3. Résultats . . . 67
2.4.3.1. Encodeur supervisé . . . 67
2.4.3.2. Encodeur auto-supervisé . . . 68
2.4.4. Discussion . . . 70
2.5. Synthèse . . . 70

2.1. Introduction

Dans ce chapitre, nous allons étudier les différentes approches et problématiques de la comparaison des lésions à l'aide de l'apprentissage automatique pour le diagnostic du mélanome. Comme vu précédemment, nous nous inspirons des recherches antérieures en dermatologie, notamment du concept du vilain petit canard (VPC) pour développer ces approches.

Pour ce faire, nous allons d'abord étudier la littérature, examiner les différents chemins de recherche déjà explorés, identifier leurs problèmes et déterminer comment nous pouvons combler les lacunes dans ce domaine. Cette étude de la littérature nous a permis d'élaborer deux idées majeures que nous regroupons dans ce chapitre. La première se base sur un apprentissage entièrement non supervisée de représentation visuelles des lésions, utilisant uniquement la différence des lésions entre elles dans l'espace de représentations pour détecter les anomalies. La deuxième se base sur une approche supervisée où les lésions d'un même patient sont directement fournies en entrée à un modèle issus des Transformers capable de comparer les images. Enfin, nous comparons les différentes approches existantes de la littérature.

Plus précisément, l'architecture générale pour comparer les images entre elles consiste en deux parties qui peuvent être distinctes ou unies : d'une part, l'encodage des images en vecteurs ; d'autre part, la comparaison de ces vecteurs dans l'espace de représentations. Cet espace de représentation correspond à l'espace dans lequel se trouvent les vecteurs que nous allons récupérer. L'important est de trouver une taille de vecteurs qui permette un bon équilibre entre capacité de représentation (grande taille) et simplicité de comparaisons (petite taille). En effet, bien que les vecteurs comportant beaucoup de caractéristiques permettent de représenter les images de manière plus précise et complexe, ils sont plus difficiles à comparer en raison de la malédiction de la dimensionnalité [52] [53]. La malédiction de la dimensionnalité réfère au phénomène par lequel l'augmentation du nombre de dimensions d'un espace entraîne une augmentation exponentielle du volume de cet espace, rendant les données disponibles de plus en plus dispersées. Cette dispersion complique la tâche des algorithmes de machine learning, car les distances entre les points deviennent moins significatives, et les modèles risquent de surajuster et de perdre en généralisation.

2.2. Revue de la littérature

La modélisation du signe du vilain petit canard a été développée de multiples façons, mais reste relativement rare dans la littérature. À ce jour, sur plus de deux millions d'articles sur le mélanome présents sur Google Scholar (un site de référencement d'articles scientifiques très utilisé en recherche), moins de 1 500 contiennent les mots-clés ugly-duckling (le concept du VPC en anglais), soit environ 0,08 %. Pourtant, les nombreuses récentes innovations dans le domaine de l'apprentissage automatique, telles que les Transformers, les approches auto-supervisées et les algorithmes de détection d'anomalies, représentent autant de nouvelles méthodes qui pourraient permettre

une détection plus précise et robuste des vilains petits canards. Ces nouvelles méthodes, capables de comparer directement plusieurs instances, permettent en effet de comparer plusieurs lésions chez un même patient. Ces éléments suggèrent qu'il reste de nombreuses pistes à explorer dans ce domaine de recherche. Dans cette section, nous passons en revue ces approches et évaluons leurs avantages et inconvénients tels que décrits dans les articles respectifs.

2.2.1. Caractérisation du VPC par des participants

Les premières utilisations du signe du vilain petit canard (VPC) pour détecter les mélanomes ont d'abord été réalisées exclusivement par des dermatologues, sans aucune automatisation. Dans un article de 2008 [54], 34 participants ont été invités à identifier des lésions cutanées atypiques en utilisant le signe du vilain petit canard pour chaque patient dans un ensemble de données. Les lésions ayant recueilli la majorité des votes des participants ont ensuite été biopsiées et analysées. Bien que l'ensemble de données soit limité (comprenant 5 mélanomes et 140 lésions bénignes), les résultats présentés sont prometteurs, avec la détection de 5 mélanomes grâce à cette technique. Cependant, la biopsie étant nécessaire pour déterminer avec certitude si une lésion est cancéreuse ou non, il est important de noter que puisque seules les lésions ayant recueilli le vote majoritaire ont été biopsiées, il n'est pas possible de déterminer si les autres lésions étaient réellement bénignes ou non.

Plus tard, en 2017, une étude très similaire avec un ensemble de données plus important (80 patients et 7 mélanomes) a également montré des résultats prometteurs [55], affirmant que "l'accès à l'analyse comparative intrapatient a réduit le potentiel de biopsie des nævi d'un facteur de 6,9", suggérant que cette voie de recherche permet de diminuer le nombre d'opérations inutiles, à savoir les exérèses sur des lésions bénignes.

Suite à ces travaux, plusieurs recherches ont progressivement commencé à intégrer des méthodes automatiques de comparaison d'images afin d'étudier ce concept de manière automatique. L'idée était maintenant d'automatiser cette tâche, puis de présenter les résultats aux dermatologues afin de leur donner un second avis lors du diagnostic. Nous allons présenter plusieurs de ces méthodes.

2.2.2. Analyse de la perte d'AutoEncodeurs

Une des premières approches entièrement automatisées du VPC se base sur les autoencodeurs [56] (AE) et leurs variantes, notamment les autoencodeurs variationnels (VAE), qui sont des réseaux de neurones entraînés à reconstruire des images après les avoir compressées dans un goulot d'étranglement (une représentation de faible dimension des images). Dans la littérature, ces modèles ont prouvé leur efficacité dans plusieurs domaines, tels que la compression de l'information et la génération d'images.

Dans le contexte de la comparaison d'images et, plus généralement, de la détection d'anomalies, ces modèles peuvent également être utilisés d'une manière plutôt ori-

ginale. Étant donné que ces modèles sont entraînés de manière auto-supervisée sur un jeu de données, ils ne peuvent recréer que des images similaires à celles qu'ils ont déjà vues. Ainsi, en analysant à quel point le modèle réussi à reconstruire une image, il est possible de prédire si l'image donnée au modèle est semblable aux images que le modèle a vu durant son entraînement, ou sinon, s'il s'agit d'une anomalie. Dans le contexte de la détection du VPC, des travaux antérieurs [57] ont exploré cette piste en entraînant des VAE uniquement sur des lésions saines, puis en développant un seuil sur la perte de reconstruction des images reconstruites. L'idée générale est que, puisque le modèle a été entraîné seulement sur des lésions saines, les images avec une perte plus élevée sont celles qui diffèrent davantage de la base d'entraînement et peuvent donc être considérées comme des anomalies. Idéalement, puisque le modèle n'a jamais vu de mélanome, il devrait afficher une perte plus élevée pour ces lésions et donc les détecter. La Figure 2.1 présente, pour un patient donné, la détection de ses lésions cutanées, ainsi que ces dernières organisées par ordre décroissant de leur atypicité selon le modèle.

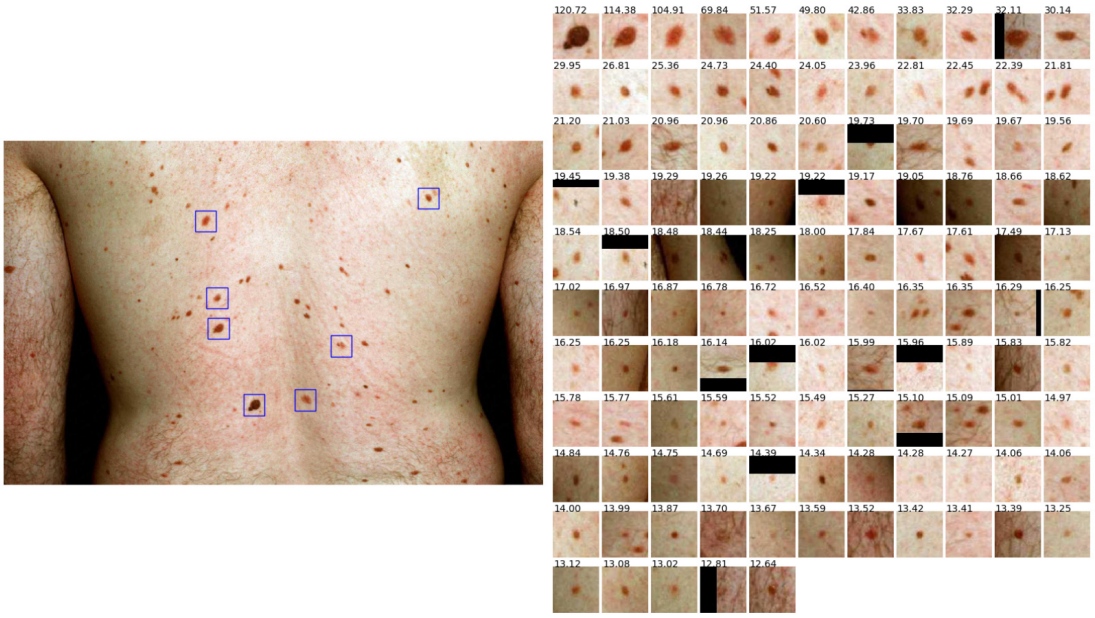

FIGURE 2.1. – **Exemple de résultat de l'approche utilisant un AutoEncodeur.** Les lésions détectées sont ensuite ordonnées par la perte qu'elles ont obtenues en passant par l'AutoEncodeur entraîné sur les lésions cutanées. *Image de l'article original [57].*

Dans ces travaux, les annotations ne sont pas "mélanome" ou "naevus", mais uniquement "anomalie" ou "lésion normale", annotées par des dermatologues. Il n'y a donc pas de résultats à proprement parler sur la classification du mélanome. Toutefois, étant donné leur annotation, le modèle prédit correctement les images 94 % du temps, avec une sensibilité à 72 % et une spécificité à 94 %.

Malgré une approche intéressante, peu de travaux futurs se sont basés sur cette

approche, dépassés par d'autres modèles plus performants et robustes.

Néanmoins, à notre connaissance, il n'y a pas d'approches mobilisant les dernières avancées en terme d'autoencodeurs, comme les beta-VAE [58] et Adversarial AutoEncoders (ALAE) [59], qui pourraient permettre de meilleures représentations, et donc de meilleures comparaisons entre lésions. En particulier, les ALAE [59] pourraient être une piste intéressante pour des travaux futurs, car permettant d'analyser les représentations apprises afin de mieux les comprendre.

2.2.3. Calcul des distances dans l'espace latent

Une autre approche consiste à utiliser les projections (ou *embeddings, plongements*) des images dermoscopiques dans l'espace latent et de les comparer directement. Dans la littérature, les travaux de Soenksen [60] ont développés une approche consistant à projeter les images d'un patient une par une dans l'espace latent, puis de comparer leur distances au centroïde des lésions. Leurs expériences sont basées sur des images à champ large, qui sont différentes des images dermoscopiques. Ces images sont en effet plutôt ressemblante à des photos classiques à laquelle la majorité de la population a accès. Le désavantage de cette base de donnée est que la résolution et les détails des images sont relativement faible par rapport aux images dermoscopiques, rendant le problème plus complexe et la robustesse des modèles moins précise. La Figure 2.2 présente l'architecture que les chercheurs ont choisi pour leurs expériences.

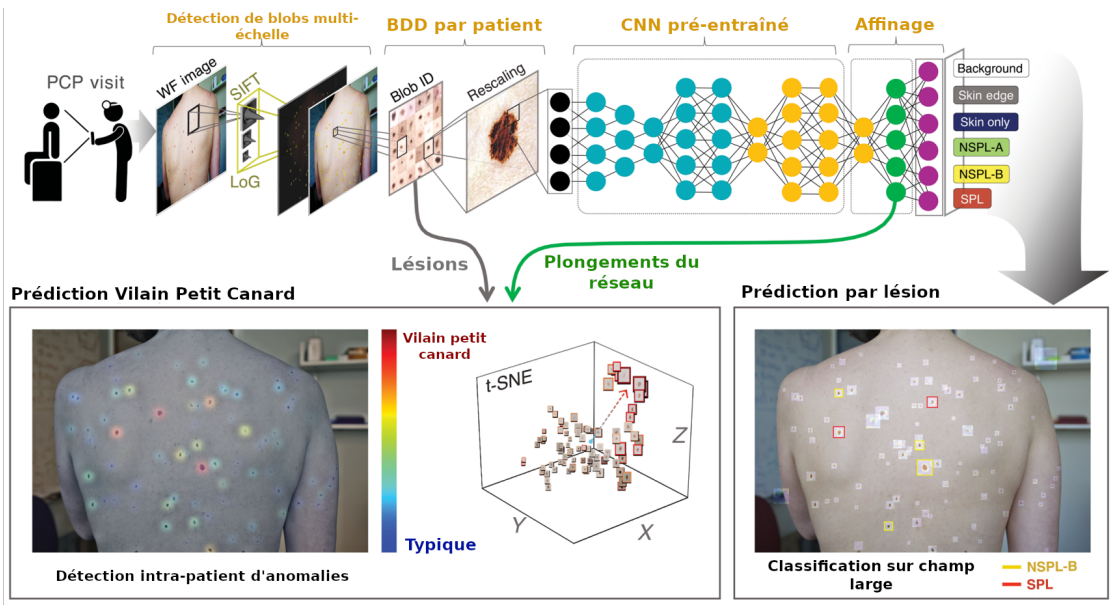

FIGURE 2.2. – **Présentation de l'architecture de l'article [60] basé sur la distance au centroïde.** Un CNN de classification est entraîné à détecter les mélanomes des naevus, puis les plongements obtenus sont comparés au centroïde du nuage de point obtenu par patient. *Image provenant de l'article original [60]*

L'approche repose sur un CNN de classification entraîné pour différencier les mélanomes des naevus. Ensuite, les auteurs extraient les caractéristiques (les sorties intermédiaires du modèle) et calculent la distance euclidienne de chaque vecteur à la moyenne. Cette dernière étape fournit un nombre pour chaque lésion, correspondant au degré d'atypicité de la lésion.

Bien que l'approche soit intéressante, la comparaison au centroïde reste une méthode très rudimentaire pour détecter les anomalies et les mélanomes. D'une part, il existe d'autres méthodes de détection d'anomalies plus avancées, telles que les algorithmes des k-plus-proches voisins [61], DBSCAN [62], ou plus récemment les modèles de mélanges gaussiens (Gaussian Mixture Models) [63, 64], les modèles à base d'énergie (Energy-based models) [65, 66], ou une combinaison de ces dernières [67]. D'autre part, comparer les lésions au centroïde revient à supposer qu'il existe une unique représentation des lésions saines, caractérisée par la moyenne des lésions, et que les lésions sont atypiques uniquement si elles diffèrent de cette moyenne. Or, ce n'est pas vrai dermatologiquement : les lésions saines se regroupent plutôt par clusters, et non autour d'un unique centre de normalité.

L'article présente néanmoins de bonnes performances, avec des résultats supérieurs à ceux obtenus par l'usage exclusif d'un CNN.

2.2.4. Approches basées sur l'attention

Enfin, ce n'est qu'à partir de 2021, avec la conception d'approches basées sur l'attention, que les performances et les architectures se sont considérablement améliorées. Les approches basées sur l'attention, initialement développées pour le traitement du langage, ont démontré leur efficacité dans le domaine de la vision. Leur méthode de fonctionnement a permis de créer des modèles capables de comparer des vecteurs de différentes images simultanément, impactant ainsi la tâche de classification des lésions cutanées. Concernant la détection du VPC, les travaux de Yu et al. [68] ont montré que les Transformers permettaient d'améliorer les performances.

Toutefois, les Transformers classiques [69], bien qu'ils montrent des résultats initiaux prometteurs [68], nécessitent énormément de ressources computationnelles et s'appuient fortement sur l'ordre des vecteurs d'entrée pour les prédictions. Dans le cas du VPC, où les données ne sont pas ordonnées et où l'ordre des vecteurs n'est pas important, il est nécessaire de ne pas utiliser leur positional embedding, normalement greffé aux vecteurs, qui permet de conserver la position. Dans notre contexte axé sur les patients, traiter les lésions cutanées au sein d'un même patient comme un ensemble plutôt que comme une séquence ordonnée pourrait s'avérer bénéfique pour la tâche de classification et pour l'efficacité de nos modèles.

Ainsi, pour développer nos approches, contrairement aux travaux précédents [68], qui utilisent une large gamme de techniques pour comparer les lésions cutanées, nous avons cherché à ce que notre architecture reste simple et adaptable, en nous concentrant uniquement sur la tâche de classification des images tout en minimisant les ressources informatiques.

Suivant cette idée, nous avons développé deux approches principales, l'une auto-

supervisée et l'autre supervisée, afin de détecter les lésions atypiques chez les patients. Nos approches se basent sur des modèles simples, efficaces et présentant un minimum d'à priori susceptibles d'influencer négativement les performances sur nos données.

2.2.5. Méthodes de détection d'anomalies

Dans nos travaux basés sur la comparaison de lésions et la détection de vilains petits canard, une partie importante de nos travaux repose sur la détection d'anomalies. Ce domaine vise à identifier les comportements atypiques ou inhabituels dans les données, qu'elles soient des images, du texte ou toute autre valeurs numérique. De nombreuses approches ont été développées pour cette tâche, couvrant une large gamme de techniques statistiques.

Les méthodes classiques de détection d'anomalies comprennent les techniques de clustering et de régression. Des algorithmes comme *k-means* et DBSCAN (Density-Based Spatial Clustering of Applications with Noise) sont utilisés pour identifier les points de données qui ne correspondent pas à un groupe ou à une distribution majoritaire. DBSCAN est particulièrement efficace pour détecter les anomalies dans les données où les groupes ne sont pas nécessairement sphériques et où il existe du bruit dans les données [62].

Parmi les méthodes basées sur l'apprentissage automatique, les Forêts d'Isolation ont émergé comme une approche puissante pour la détection d'anomalies dans de grandes quantités de données. Cet algorithme repose sur la construction d'arbres de décision pour isoler les points de données. Les points qui sont facilement isolés par ces arbres sont considérés comme des anomalies. L'efficacité de ces forêts réside dans leur capacité à traiter des ensembles de données volumineux avec une faible complexité computationnelle [70].

Une autre méthode populaire est l'utilisation des autoencodeurs, comme décrit plus haut, en particulier dans les applications impliquant des données de grande dimension. Les autoencodeurs sont des réseaux neuronaux entraînés pour reconstruire l'entrée à partir d'une représentation comprimée. Les anomalies sont identifiées en détectant des erreurs de reconstruction importantes, ce qui suggère que le modèle n'a pas été capable de reconstruire correctement des points de données rares ou inhabituels [71].

2.3. Détection non supervisée des mélanomes

2.3.1. Introduction

La première approche que nous avons développée consiste en la détection entièrement non supervisée des mélanomes en utilisant le concept du VPC. Nous avons voulu vérifier à quel point l'information du VPC avait la capacité de discriminer les lésions saines des anomalies. Contrairement aux autres approches de détection d'anomalies, qui se basent sur l'apprentissage exclusif des images saines de toute la base

d'entraînement, notre approche diffère fondamentalement en définissant les anomalies par patient, et non pour l'ensemble de la base de données. Cette particularité rend ce concept unique par rapport aux autres problèmes existants en apprentissage automatique et en vision par ordinateur, car l'anomalie n'en est pas une pour l'ensemble de la base de données, mais uniquement pour un sous-ensemble de celle-ci, à savoir le patient. Cette différence fondamentale nécessite un changement radical dans les architectures actuelles de détection d'anomalies. Premièrement, le faible nombre d'images par patient (le nombre de lésions, d'une dizaine à une centaine) est insuffisant pour entraîner un modèle à partir de zéro sur ces images. Ainsi, l'idée générale de notre approche a été de créer une méthode permettant de comparer les images dont les représentations vectorielles auraient été apprises en amont, et de manière non supervisée, afin de ne pas inclure le biais de classification que l'on aurait pu avoir en utilisant des CNN de classification.

2.3.2. Méthode

2.3.2.1. Architecture globale

Notre architecture nécessite deux étapes importantes : l'encodage, puis la comparaison. La première étape d'encodage nous permet de projeter les images dans un espace plus petit, appelé espace latent, dans lequel les images projetées, que l'on appelle dans cet espace les *embeddings* (ou également *plongements*, que l'on utilisera de manière interchangeable), peuvent être comparées aisément. En effet, sans réduction de l'espace de recherche, nous faisons face au fléau de la dimensionnalité, qui est un problème important en science des données, mettant en avant la difficulté de comparer et de calculer les distances lorsque les données se situent dans un espace doté de trop nombreuses dimensions. Notre encodage permet ainsi de réduire ce problème en utilisant un espace vectoriel plus petit. Ensuite, nous comparons, par patient, les images. Nous avons choisi une approche ayant du sens d'un point de vue dermatologique, en utilisant une variante des k-plus proches voisins. De la sorte, les lésions regroupées en clusters seront considérées comme normales, lorsque les lésions très atypique et différentes de toutes les autres seront considérées comme des anomalies. Cet algorithme permet ainsi la prédiction d'un "taux d'atypicité" (que nous avons appelé Ugly-Duckling Index, UDI, en anglais), qui sera utilisé pour la prédiction. La partie finale consiste alors à choisir un seuil du UDI pour lequel un taux trop élevé d'atypicité est prédit comme une anomalie, et un taux faible comme une lésion saine. Notre architecture est présentée à la Figure 2.3.

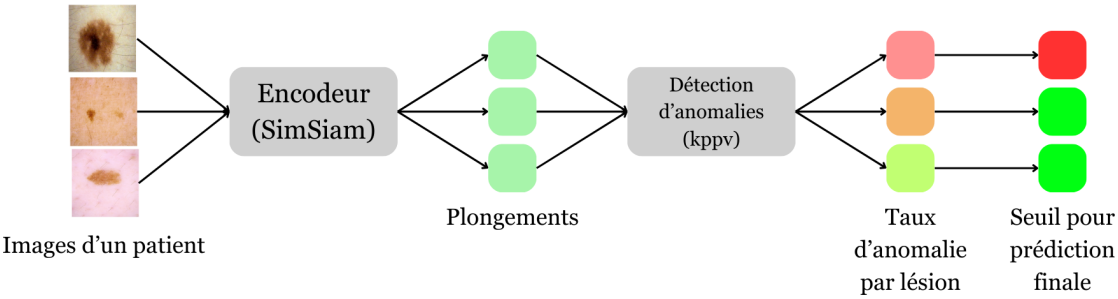

FIGURE 2.3. – **Architecture de notre approche**. L'encodeur transforme les images en vecteurs, puis ces derniers sont comparés à l'aide de l'algorithme des k-plus-proches-voisins. Enfin, le seuil optimal est choisi empiriquement sur la base de validation.

2.3.2.2. L'architecture SimSiam comme encodeur

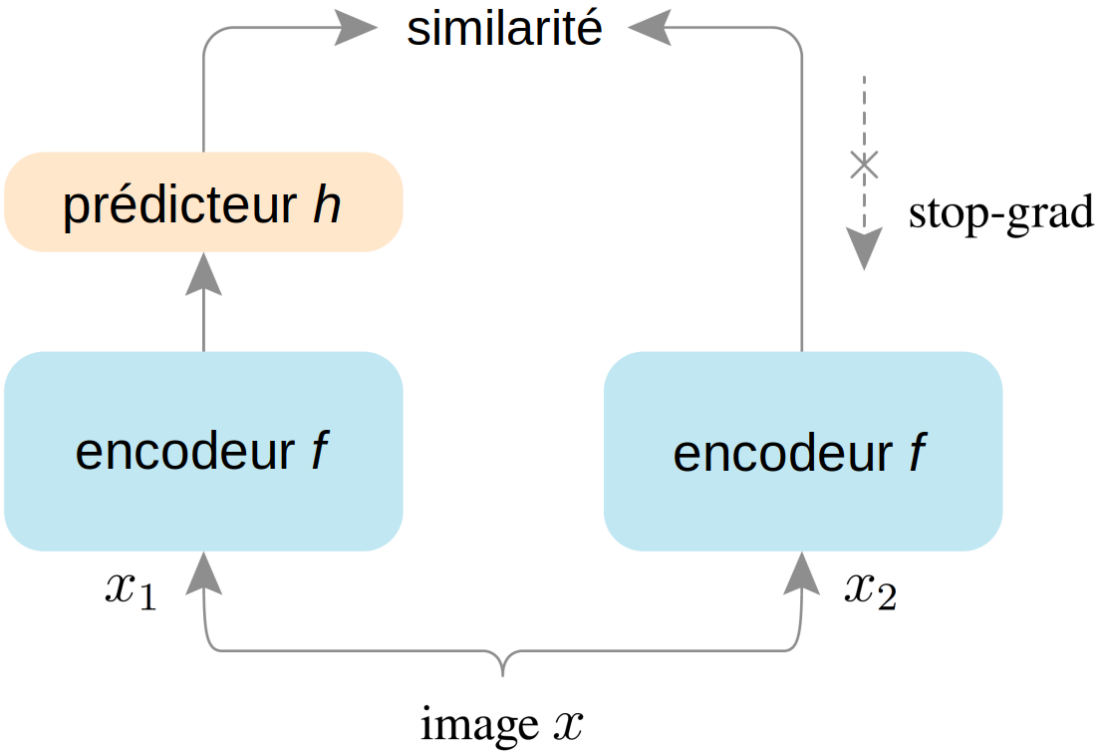

FIGURE 2.4. – **Diagramme de l'architecture SimSiam [72].** *Image provenant de l'article original.*

Notre premier module est donc la compression des images dans l'espace latent. Pour ce faire, nous avons utilisé le CNN ResNet-50 [36]. Cependant, comme les CNN classiques sont entraînés à l'aide d'annotations, nous ne pouvons pas utiliser ces méthodes d'entraînement traditionnelles. Pour surmonter cette contrainte, nous

avons adopté une approche auto-supervisée, qui permet d'entraîner le modèle à générer des embeddings pertinents sans nécessiter d'annotations. L'approche que nous avons choisie est SimSiam (voir Figure 2.4), décrite dans l'article éponyme par Xinlei Chen et al. en 2020 [72]. Cette méthode se base sur un entraînement dit "siamois", signifiant que le modèle, à chaque étape de l'entraînement, prend en entrée deux entrées et calcule la distance entre elles pour déterminer la perte. Cette approche existe sous de multiples formes et variantes, telles que les architectures SwaV [73], SimCLR [74] et BYOL [75], que nous décrirons plus tard. Dans le cas de SimSiam, les deux images fournies en entrée à chaque étape sont en réalité la même image augmentée différemment, ce qui permet au modèle d'apprendre les caractéristiques fondamentales de l'image sans nécessiter d'annotations.

Plus précisément, SimSiam a la qualité d'être très économe en ressources, et n'a également ni besoin d'annotations, ni de paires négatives. Les paires négatives sont couramment utilisées dans la littérature pour de nombreuses méthodes contrastives, consistant à donner au modèle deux images différentes, puis à lui faire apprendre à éloigner ces deux images dans l'espace vectoriel (sous la forme d'une fonction de perte qui utilise la distance entre les deux embeddings).

Dans le cas de SimSiam, cette opération n'est pas réalisée, ce qui pourrait laisser croire que les images se rapprochent continuellement dans l'espace vectoriel. En réalité, différentes méthodes sont mises en œuvre pour empêcher une telle dégénérescence, notamment en se basant sur deux opérations clés : l'arrêt du gradient sur une branche du modèle et l'ajout d'un "prédicteur" sur une autre branche. Étant donné que l'encodeur n'est jamais optimisé seul directement et que les images des deux branches sont toujours différentes en raison de la data augmentation, le modèle reste capable de générer des embeddings variés et espacés.

Pour nos expériences, nous avons conservé les hyperparamètres par défaut, décrits dans l'article original. Nous avons seulement réduit la taille des embeddings, en passant de 2048 à 100, afin de simplifier la tâche suivante de comparaison, basée sur la distance des vecteurs. Cette réduction de dimension a été réalisée avec l'algorithme UMAP [76]. Enfin, les valeurs des vecteurs sont normalisées sur la base d'entraînement. L'entraînement est réalisé sur 150 époques, en utilisant la fonction de perte de similarité cosinus, l'optimiseur SGD ainsi qu'un taux d'apprentissage de 0,05.

Avant d'entrer dans le modèle, les images sont redimensionnées à 224x224, qui est la taille de base pour ResNet-50. Les méthodes de data augmentation sont restées les mêmes que pour l'article SimSiam, à savoir des techniques aléatoires de redimensionnement, de rognage, de flou gaussien, d'image miroir et de modifications de couleurs (voir Figure 2.5). Le modèle semble avoir convergé vers la 150ème époque.

2.3.2.3. Une variante des k-plus-proches voisins pour comparer les images

Une fois les images encodées, nous pouvons les comparer pour détecter les anomalies. Ce problème est généralement simple à traiter, mais l'organisation de nos données rend le problème plus complexe. En effet, comme mentionné précédemment, la ma-

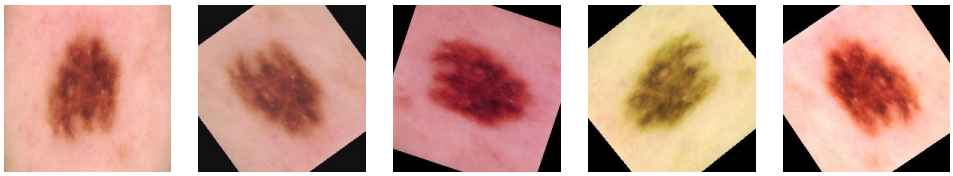

FIGURE 2.5. – **Exemples de la data augmentation utilisée.** L'image originale est affichée tout à gauche, suivie de la même image transformée aléatoirement de 4 manière différentes.

jorité des problèmes de détection d'anomalies reposent sur des ensembles complets de données, comportant au moins des centaines d'instances, voire plusieurs millions. Les modèles de détection d'anomalies sont également généralement créés de sorte à détecter les anomalies dans un ensemble de données bien défini comportant de nombreuses images. Or, le problème du VPC comporte une subtilité très importante : une lésion atypique pour un patient ne l'est pas forcément pour un autre patient. Nos sous-ensembles contiennent alors finalement au mieux une centaine de lésions, au pire moins d'une dizaine. Détecter une anomalie parmi dix instances devient alors un problème très difficile, car cela comporte beaucoup d'incertitudes. De plus, utiliser des méthodes d'apprentissage serait encore bien plus difficile vu la faible quantité de données. Nous avons donc cherché à développer un algorithme qui puisse pallier ces problèmes, qui ne requiert pas énormément de données et qui puisse s'adapter à chaque patient. Finalement, l'approche que nous avons développée est une variante de l'algorithme des k-plus-proches voisins : pour chaque lésion d'un patient, nous la comparons avec ses k-plus-proches voisins (kppv) en calculant sa distance moyenne à ses kppv. Cela nous donne une distance que nous comparons avec les autres distances obtenues avec les autres lésions. De manière formelle, l'algorithme vise à calculer l'UDI (indice d'atypicité) de chaque lésion que nous allons décrire ici. Nous calculons tout d'abord la distance moyenne des lésions à leurs k voisins avec l'équation 2.1.

$$D_i = \frac{\sum_{j \in K_i} d(v_i, v_j)}{k} \tag{2.1}$$

Avec v_i le vecteur de caractéristiques de l'image i, k le nombre de voisins choisi, K_i l'ensemble des k voisins de i, et d la fonction de distance (dans notre cas la distance cosine). Nous avons donc pour chaque lésion sa distance moyenne à ses k voisins. Ces valeurs obtenues sont ensuite divisées par leur moyenne (dénotée par $\overline{D}$) au sein du patient afin d'obtenir le score d'atypicité final (Equation 2.2).

$$UDI_i = \frac{D_i}{\overline{D}} \tag{2.2}$$

Enfin, nous normalisons les valeurs sur toute la base de données, afin que les valeurs aient une moyenne nulle et une variance unitaire.

Nous notons d'ailleurs que l'usage de la distance cosinus convient parfaitement à l'entraînement de SimSiam, dont la fonction de perte, la similarité cosinus, est liée à

notre distance par cette simple équation :

$$cosine_distance = 1 - cosine_similarity$$

La figure 2.6 présente une simple visualisation en 2 dimension d'un exemple de résultats pour un patient hypothétique.

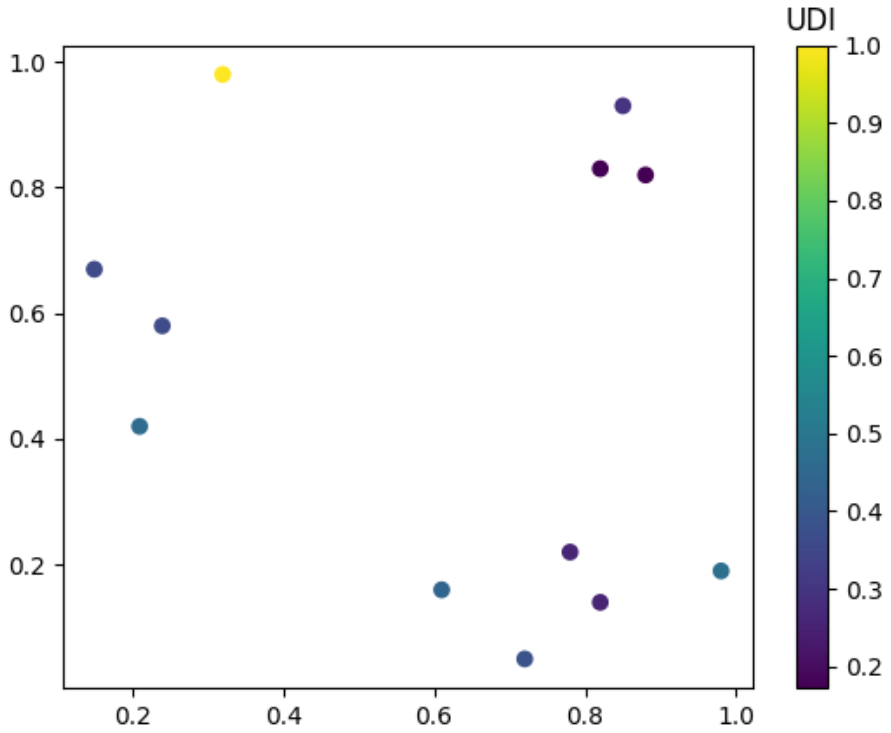

FIGURE 2.6. – **Visualisation de notre algorithme de détection d'anomalies.** Dans cet exemple, $k = 1$ est utilisé sur un nuage de points en 2 dimensions. Chaque point représente le vecteur de caractéristiques d'une lésion tandis que sa couleur représente son UDI.

Dans les cas où les patients n'ont pas suffisamment de lésions (moins de lésions que l'hyperparamètre k), nous calculons la distance des lésions avec le centroïde des lésions, similairement à l'approche utilisée par Soenksen [60].

2.3.3. Résultats

L'évaluation de notre modèle sur la base de test est réalisée en comparant le taux d'atypicité prédit avec les annotations des images (mélanome ou naevus). Dans cette section, nous analysons les performances de notre modèle, d'abord en utilisant plusieurs valeurs de k sur la base de validation, puis en utilisant le k optimal sur la base

de test.

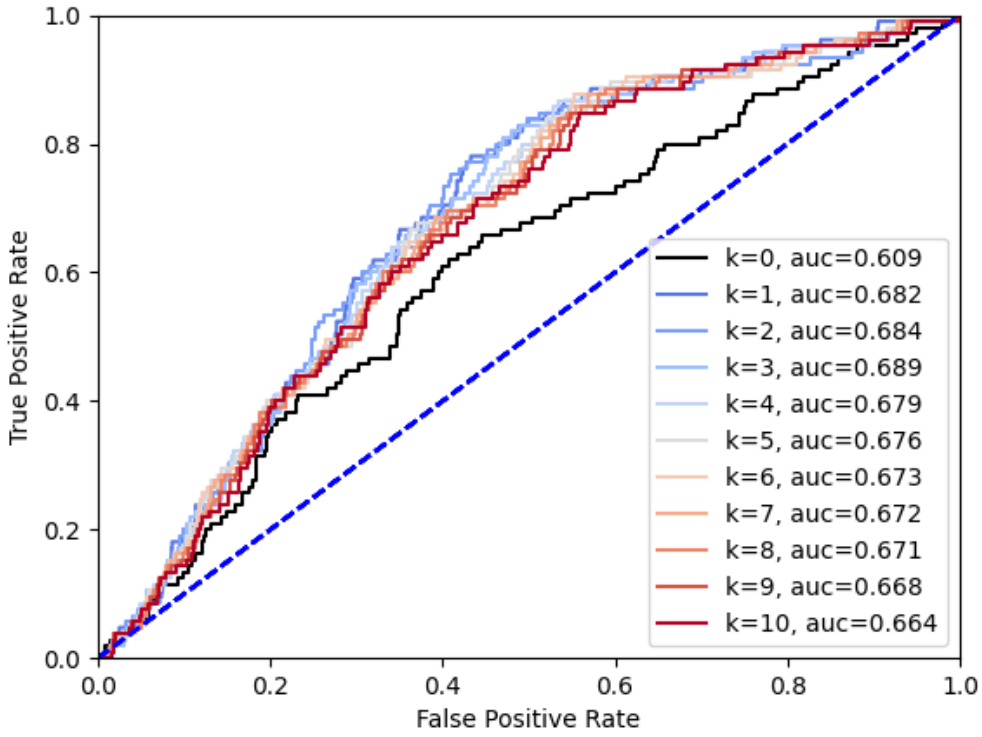

FIGURE 2.7. – **Performance sur la détection de mélanome en fonction de k. k=0 compare les lésions avec le centroid.**

L'évaluation a d'abord été axée sur la détermination de la valeur optimale pour le paramètre k sur les données de validation. Notre étude a révélé une tendance claire : l'augmentation de la valeur de k entraînait une amélioration des performances jusqu'à atteindre un pic à $k = 3$, après quoi la performance a progressivement diminué. Notamment, lorsque nous utilisons la méthode de la distance au centroïde, l'aire sous la courbe ROC (AUC) atteint sa valeur la plus basse, indiquant que notre algorithme des k plus proches voisins permet une comparaison plus approfondie des lésions que les approches précédentes. Nous supposons que les lésions atypiques, étant typiquement isolées, peuvent être adéquatement évaluées en les comparant à leurs trois voisins les plus proches pour déterminer leur écart par rapport aux autres lésions. D'un point de vue dermatologique, cela fait sens, car cette approche représente plusieurs clusters de normalité, au lieu d'un seul. Ainsi, les lésions présentes sur différentes parties du corps et celles présentant des motifs similaires ne sont alors plus considérées comme des anomalies, mais comme des ensembles sains, bien que différents.

Suite à l'identification de la valeur optimale pour k sur les données de validation, nous avons évalué la performance de notre modèle sur l'ensemble de test en utilisant la valeur sélectionnée de $k = 3$. Étonnamment, l'AUC correspondante est de 0,719,

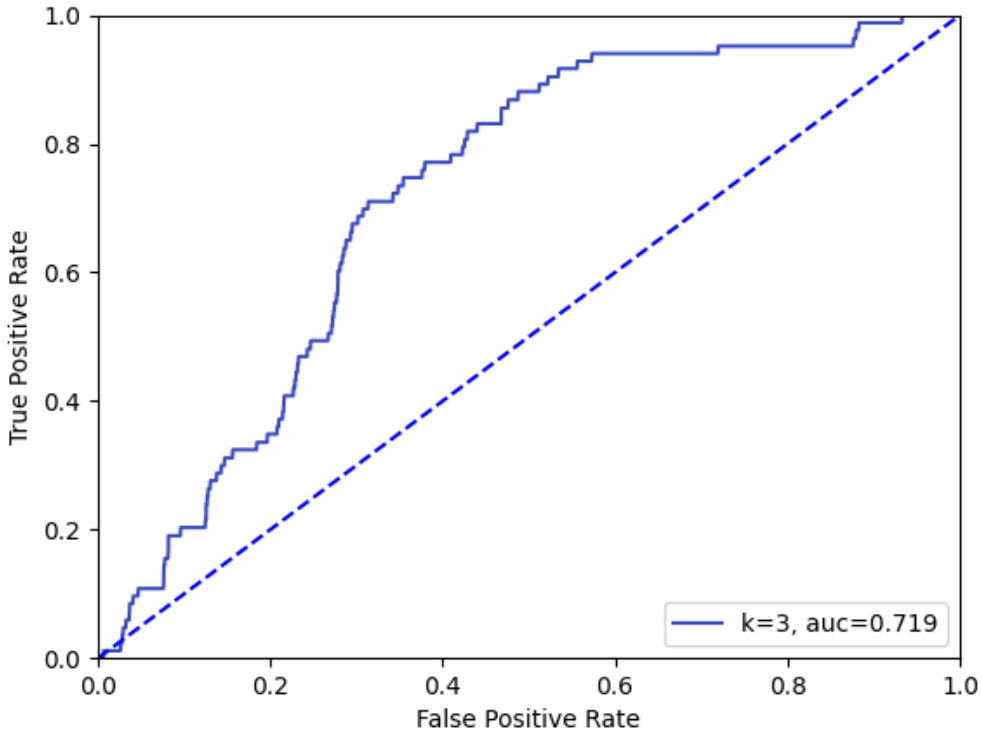

FIGURE 2.8. – **Résultats sur la base de test après avoir choisi k=3 grâce à la base de validation.**

indiquant une performance encore supérieure par rapport aux données de validation. La précision équilibrée maximale atteinte par la sélection du seuil est de 0,70, correspondant à une sensibilité de 0,77 et une spécificité de 0,62. Bien que ces résultats restent inférieurs à la performance des modèles qui exploitent les annotations de référence, ils sont particulièrement prometteurs, surtout si l'on prend en compte que les annotations sont entièrement omises de l'ensemble du pipeline. En continuant cette piste de recherche, il est fort probable que de nouvelles approches similaires pourront être développées, en améliorant la partie encodage et comparaison à l'aide d'autres architectures et algorithmes de comparaison.

2.3.4. Conclusion

Cette première étude présente des résultats convaincants en exploitant un modèle siamois entièrement auto-supervisée sur des images dermoscopiques. Malgré la difficulté inhérente de cette tâche, notre architecture démontre une compétence dans la détection du mélanome, évitant le recours à des données étiquetées et s'appuyant uniquement sur les comparaisons inter-lésions au sein des patients. Cette approche novatrice offre plusieurs contributions significatives. D'une part, elle permet de dé-

tecter les anomalies par patient, sans jamais avoir accès à aucune annotation, ce qui est unique dans a littérature du VPC. De plus, contrairement à la majorité des travaux qui se concentrent sur les anomalies à l'échelle de bases de données entières, notre approche évalue les anomalies au sein de patients présentant un faible nombre de lésions, généralement inférieur à 100. Notre recherche montre également que le concept du VPC permet, de manière indépendante et dans une certaine mesure, la détection des mélanomes. Bien que tous les VPC ne soient pas des mélanomes, ils composent une part non négligeable de ces derniers, que notre approche est capable d'identifier. Ces avancées ouvrent la voie à des perspectives prometteuses, notamment en intégrant ces informations dans des modèles de classification, en explorant d'autres modèles d'apprentissage de représentation visuelle, ainsi qu'en ajoutant une part d'interprétabilité des vecteurs appris. Ces développements pourraient avoir un impact substantiel sur des applications plus larges, notamment dans le domaine du diagnostic de divers types de cancers et de maladies, et permettant à terme une implantation dans le milieu clinique afin d'accélérer et d'améliorer les diagnostics.

2.4. Détection supervisée du mélanome par comparaison de lésions cutanées par patient

2.4.1. Introduction

Dans notre précédente approche, nous avons analysé les performances de la détection de lésions atypiques de manière entièrement non supervisée. Si cela nous a permis d'obtenir de bons résultats et de développer une approche originale, nous avons ensuite voulu ajouter les annotations dans l'entraînement de notre modèle afin de vérifier les gains potentiels de performance, cette fois-ci en utilisant les annotations durant l'entraînement. Bien que cela constitue un changement de paradigme par rapport à l'approche précédente, les découvertes que les expériences antérieures nous ont apportées peuvent permettre d'améliorer les méthodes supervisées. Ainsi, l'idée de notre seconde méthode est de réaliser une classification, entraînée à l'aide des annotations, tout en comparant les images entre elles au sein d'un même patient. Le diagnostic est ainsi enrichi par la tâche de comparaison des images, contrairement aux classifications classiques prenant en entrée seulement une unique image par itération. Pour ce faire, nous avons conservé notre architecture globale, composée d'un encodeur et d'un détecteur d'anomalies. Ce travail, bien que ressemblant au précédent, diffère de plusieurs manières. Premièrement, il ne s'agit plus concrètement de détection d'anomalies, mais plutôt de comparaison d'images, ce qui permet au modèle de connaître, de manière globale, la structure des lésions du patient. Deuxièmement, la partie supervisée permet de comparer les performances avec et sans comparaison d'images, ce qui permet de vérifier si cette dernière améliore les performances.

Notre architecture, dénommée ReSeT, (illustrée dans la Figure 2.9) repose sur un encodeur en conjonction avec un Set-Transformer [77], l'un réalisant l'encodage des images, l'autre la comparaison et classification. L'encodeur, entraînable soit de

manière supervisée, soit auto-supervisée, génère les embeddings des images qui seront utilisés à la fois pour comparer les images entre elles que pour les classer entre les catégories mélanome et naevus. Nous utilisons encore une fois les modèles auto-supervisés en tant qu'encodeur, car ces modèles, en s'abstenant d'utiliser des annotations de vérité terrain, garantissent que les embeddings extraits restent impartiaux par rapport à la tâche de classification d'origine. Cela est particulièrement crucial pour obtenir une représentation plus diversifiée des images, générant des embeddings représentant des caractéristiques qui ne seraient pas nécessairement apprises par un modèle supervisé. En effet, l'usage de plusieurs modèles travaillant en même temps pour une même tâche finale est une approche très utilisée dans les méthodes ensemblistes (méthodes qui combinent de nombreux modèles différents). Bien plus que la simple utilisation de plusieurs modèles, l'idée est de développer des représentations diverses des images. C'est cette diversité dans les approches qui va permettre, idéalement, d'avoir des représentations visuelles variées, capables d'analyser de nombreuses caractéristiques différentes dans les images. La suite de notre approche se compose d'un Set-Transformer, utilisé pour comparer ces embeddings et générer des prédictions diagnostiques. Ce modèle particulier, que nous décrivons plus en détail dans la partie suivante, présente plusieurs avantages nécessaires à notre approche, comme son mode de fonctionnement équivariant à la permutation, garantissant des prédictions cohérentes indépendamment de l'ordre des lésions, ce qui est le cas dans notre tâche. Il y a également ses calculs, optimisés pour réduire la charge computationnelle en facilitant les calculs par des points d'inductions.

2.4.2. Méthode

2.4.2.1. ReSeT

Nous présentons ReSeT, une approche pour comparer les lésions cutanées. Notre modèle se compose de deux modules principaux : un encodeur, et un Set-Transformer, entraînés indépendamment. Tout d'abord, l'encodeur est entraîné à dériver des embeddings d'images, essentiels pour les opérations ultérieures dans le Set-Transformer. Ensuite, le Set-Transformer est entraîné sur la tâche de classification, prenant en entrée les embeddings générés. Nous utilisons les mêmes ensembles d'entraînement, de validation et de test que pour l'encodeur.

Le Set-Transformer produit une sortie pour toutes les entrées, reflétant le contexte des lésions du patient. De plus, nous augmentons l'architecture originale du Set-Transformer en introduisant une couche résiduelle finale qui relie les embeddings à la sortie correspondante. Cette modification implique de concaténer ces sorties avec leurs vecteurs d'entrée respectifs, plutôt que de se fier uniquement aux sorties du Set-Transformer. Enfin, une seule couche dense est utilisée pour faire les prédictions.

Les couches résiduelles garantissent que la couche finale du réseau assimile à la fois le vecteur d'origine et les informations contextuelles du Set-Transformer. Tandis que l'encodeur apprend à mapper les images dans l'espace latent, le Set-Transformer apprend à faire des prédictions non seulement en utilisant les images individuelles,

mais aussi en tirant parti des informations supplémentaires liées à toutes les autres lésions. Cette approche permet au modèle d'utiliser soit les informations contextuelles, soit simplement de s'appuyer sur le plongement d'image, selon le patient, car certains patients peuvent ne pas avoir beaucoup de lésions cutanées.

Cette méthode adresse également les problèmes de la disparition des gradients grâce aux connexions résiduelles tout en maintenant un équilibre entre les caractéristiques individuelles des images et l'information du contexte du patient.

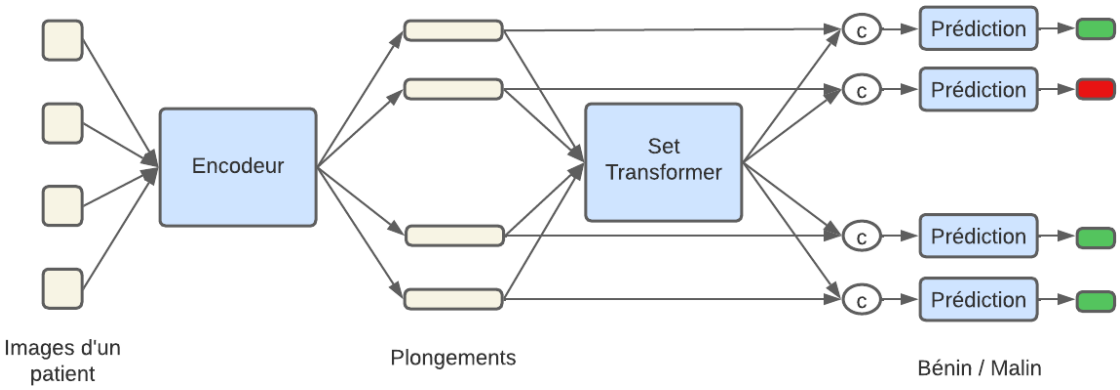

FIGURE 2.9. – **Diagramme de l'architecture ReSeT**. L'encodeur prend en entrée les images une par une, puis le Set-Transformer récupère tous les vecteurs d'un coup pour les comparer. Enfin, les sorties du Set-Transformer sont concaténées avec les vecteurs issus de l'encodeur afin de générer la prédiction finale bénin/malin.

2.4.2.2. Encodeur

Nous avons utilisé ResNet-50 [36] comme modèle de base pour les architectures chargées d'encoder les images. Étant donné que l'architecture originale offre un espace latent avec plus de 1 000 caractéristiques, ce qui est trop élevé dans notre contexte, nous avons ajouté une couche dense avec une taille de sortie de 200, visant à générer des plongements concis et comparables plutôt que des vecteurs de haute dimension. Ensuite, nous avons développé deux approches différentes pour l'encodage des images afin de comparer leurs résultats. Nous avons d'une part l'approche supervisée, d'autre part l'approche non-supervisée.

Dans notre configuration d'entraînement supervisé, nous avons entraîné le modèle ResNet-50 sur une tâche de classification distinguant entre le mélanome et le nævus pendant 200 époques, en utilisant l'optimiseur Adam, un taux d'apprentissage fixé à 0,001 et une taille de batch de 64. La sélection du modèle était basée sur l'aire sous la courbe ROC (AUC) à partir de l'ensemble de validation. Après l'entraînement, nous avons extrait les plongements d'images de l'avant dernière couche de taille 200.

Puis, nous avons également exploré les performances de notre architecture en utilisant des modèles auto-supervisés comme encodeurs. En exploitant les données non étiquetées grâce à l'apprentissage auto-supervisé, nous pouvons capturer des

structures de données sous-jacentes au-delà des simples étiquettes de classification. La plupart de ces méthodes auto-supervisées permettent de générer un espace latent où la proximité indique la similarité des images, ce qui offre des avantages dans les tâches de classification, de comparaison ainsi que dans de potentielles tâches de visualisations utiles dans les environnements cliniques.

Nous avons comparé les principales architectures auto-supervisées actuelles telles que BYOL [75], MoCo [78], SimCLR [74], SwaV [73], et SimSiam [72]. Chaque modèle a été entrainé sur 200 époques d'entraînement avec une taille de batch de 128, avec des taux d'apprentissage par défaut spécifiés dans les articles associés, et en maintenant une taille de sortie normalisée de 200. Ces modèles ont été initialisés de manière aléatoire, sans l'utilisation des poids ImageNet, évaluant ainsi la capacité de notre architecture à discerner le mélanome du nævus sans connaissance préalable. En dehors de cela, tous les autres hyperparamètres ont été maintenus identiques à ceux décrits dans leurs articles associés.

2.4.2.3. Comparaison d'images à base de SetTransformers

Les Set-Transformers sont bien adaptés à notre tâche de comparaison des lésions cutanées en raison de leur efficacité dans le traitement des ensembles de données non ordonnés. L'article original [77] introduit des opérations telles que SAB (Set-Attention Block), ISAB (Induced SAB), Multihead Attention Block (MAB) et Pooling by Multihead Attention (PMA), permettant l'invariance par permutation et l'optimisation de la complexité computationnelle en utilisant des points d'induction, traitant ainsi efficacement les données non ordonnées tout en minimisant les coûts computationnels (voir Figure 2.10). L'idée générale de cette architecture est que le modèle ne va pas utiliser de *positional embedding* comme les Transformers classiques, afin de garder le modèle invariant à la permutation. De plus, les points d'inductions, à travers les fonctions ISAB vont permettre la représentation des images sous forme de clusters de données regroupées dans les points d'inductions, limitant également les calculs réalisés.

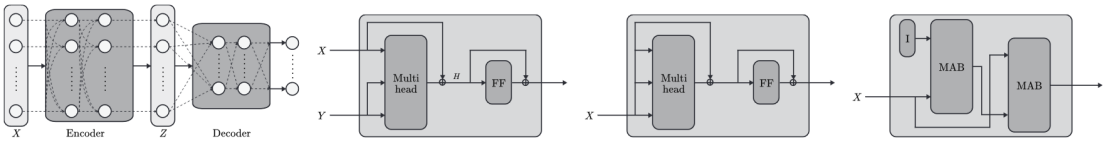

FIGURE 2.10. – **Diagramme de l'architecture des Set-Transformers et ses différentes opérations.**

Le Set-Transformer que nous avons choisi adopte une structure encodeur-décodeur, similaire à l'implémentation originale. L'encodeur comprend deux blocs ISAB, tandis que le décodeur comprend une couche PMA, deux blocs SAB et une couche dense finale. Nous avons fixé la dimension cachée à une taille de 128 et avons utilisé 8 têtes d'attention.

Pendant l'entraînement, le Set-Transformer traite la liste d'images de chaque patient (représentées sous forme de vecteurs), générant des vecteurs d'informations contextuelles de taille 2 pour chaque image. Ces sorties sont concaténées avec les vecteurs d'entrée d'origine et alimentées dans la couche de prédiction finale. Avec une petite sortie de taille 2 pour les images, nous nous assurons que les vecteurs générés représentent des informations utiles pour la prédiction.

L'entraînement s'est étendu sur 200 époques, en utilisant une perte d'entropie croisée et l'optimiseur SGD. Pour éviter le surajustement, l'entraînement s'est arrêté lorsque l'AUC de validation n'a pas augmenté après 15 époques. Chaque étape d'entraînement contenait 64 patients, avec 20 vecteurs aléatoires sélectionnés par patient, dépassant la moyenne de 16 lésions par patient dans le jeu de données ISIC 2020. Pour les patients avec moins de lésions, des vecteurs nuls ont été utilisés comme rembourrage. La validation et les tests ont traité les patients avec un excès d'images en les soumettant plusieurs fois au réseau pour la prédiction.

2.4.3. Résultats

Dans cette section, nous analysons les résultats obtenus sur notre tâche de classification.

Pour chaque architecture, nous avons sélectionné le modèle avec le plus haut AUC sur l'ensemble de validation, et nous avons utilisé une méthodologie de recherche par grille (ou grid search en anglais) sur l'ensemble de validation pour trouver le seuil de classification optimal utilisé pour le calcul de la précision équilibrée, de la sensibilité et de la spécificité. Ces modèles optimisés, ainsi que leurs seuils respectifs, ont ensuite été utilisés pour la prédiction finale sur l'ensemble de test.

De plus, pour évaluer de manière exhaustive les performances des architectures, nous avons entraîné cinq Set-Transformers distincts pour chaque architecture puis avons calculé leur moyenne et leur écart-type. Cette méthode nous a permis d'analyser et de comparer la stabilité et les performances moyennes entre les architectures pour chaque métrique.

Nous comparons les résultats de notre approche avec les résultats obtenus sur un CNN classique de classification, ainsi qu'une variante de notre approche sans la couche résiduelle. Cette dernière obtient directement en entrée les vecteurs, et réalise le diagnostic directement en sortie du Set-Transformer.

2.4.3.1. Encodeur supervisé

Nous commençons par évaluer les performances de l'approche supervisée. Cela implique d'entraîner un ConvNet de classification sur la tâche binaire de détection du mélanome et du nævus et de l'utiliser comme encodeur de l'architecture ReSeT. Les résultats sont détaillés dans les Tableaux 2.1 et 2.2.

Le CNN de classification a bien fonctionné, obtenant une AUC de 0,905, en accord avec les performances attendues pour cette tâche. De manière intéressante, l'usage direct des caractéristiques dans un Set-Transformer sans couche résiduelle a entraîné

Architecture supervisée	AUC	Précision équilibrée
CNN de classification	0.905 ± 0.01	0.707 ± 0.01
Set-Transformer (sans résiduel)	0.801 ± 0.01	0.690 ± 0.02
ReSeT	$\mathbf{0.924 \pm 0.01}$	$\mathbf{0.831 \pm 0.02}$

TABLEAU 2.1. – **AUC et précision équilibrée de la classification sur la base de test.** Nous comparons les performances entre un CNN de classification classique, un SetTransformer entrainé sur les caractéristiques du CNN ainsi que les performances de ReSeT sur les caractéristiques du CNN.

Architecture supervisée	Sensitivité	Spécificité
CNN de classification	0.775 ± 0.06	0.705 ± 0.06
Set-Transformer (sans résiduel)	0.798 ± 0.09	0.583 ± 0.12
ReSeT	$\mathbf{0.874 \pm 0.05}$	$\mathbf{0.788 \pm 0.09}$

TABLEAU 2.2. – **Sensitivité et spécificité de la classification sur la base de test.**

une diminution des performances, avec une AUC de 0,801. Cette baisse pourrait résulter des difficultés à associer chaque sortie à son entrée d'image correspondante. Enfin, notre modèle ReSeT, qui intègre des connexions résiduelles avec les caractéristiques du CNN, a démontré une amélioration significative des performances, atteignant une AUC de 0,924. Les meilleures performances obtenues résultent alors de l'utilisation de couches résiduelles ainsi que de la comparaison d'images. Ces deux éléments clés on permit au Set-Transformer d'obtenir de très bon résultats pour la détection du mélanome.

Ces conclusions soulignent le potentiel de l'utilisation des Set-Transformers dans le diagnostic dermatologique, présentant une voie prometteuse pour les cliniciens et les chercheurs. Les performances supérieures de l'approche combinée mettent en évidence les limitations à prendre en compte lors de l'usage des Transformers classiques et des Set-Transformers.

2.4.3.2. Encodeur auto-supervisé

Architecture Auto-supervisée	AUC	Précision équilibrée
BYOL	0.681 ± 0.01	0.635 ± 0.02
MoCo	$\mathbf{0.768 \pm 0.02}$	$\mathbf{0.708 \pm 0.02}$
SimCLR	0.764 ± 0.01	0.683 ± 0.02
SimSiam	0.748 ± 0.01	0.702 ± 0.01
SwaV	0.617 ± 0.01	0.569 ± 0.03

TABLEAU 2.3. – Résultats de la tâche de classification sur le jeu de test utilisant des architectures auto-supervisées pour l'entraînement du CNN.

Architecture Auto-supervisée	Sensibilité	Spécificité
BYOL	0.694 ± 0.04	0.576 ± 0.04
MoCo	0.730 ± 0.09	**0.686 ± 0.06**
SimCLR	0.780 ± 0.02	0.587 ± 0.02
SimSiam	**0.849 ± 0.01**	0.555 ± 0.02
SwaV	0.602 ± 0.16	0.536 ± 0.12

TABLEAU 2.4. – **Résultats de la tâche de classification sur le jeu de test utilisant des architectures auto-supervisées pour l'entraînement du CNN.**

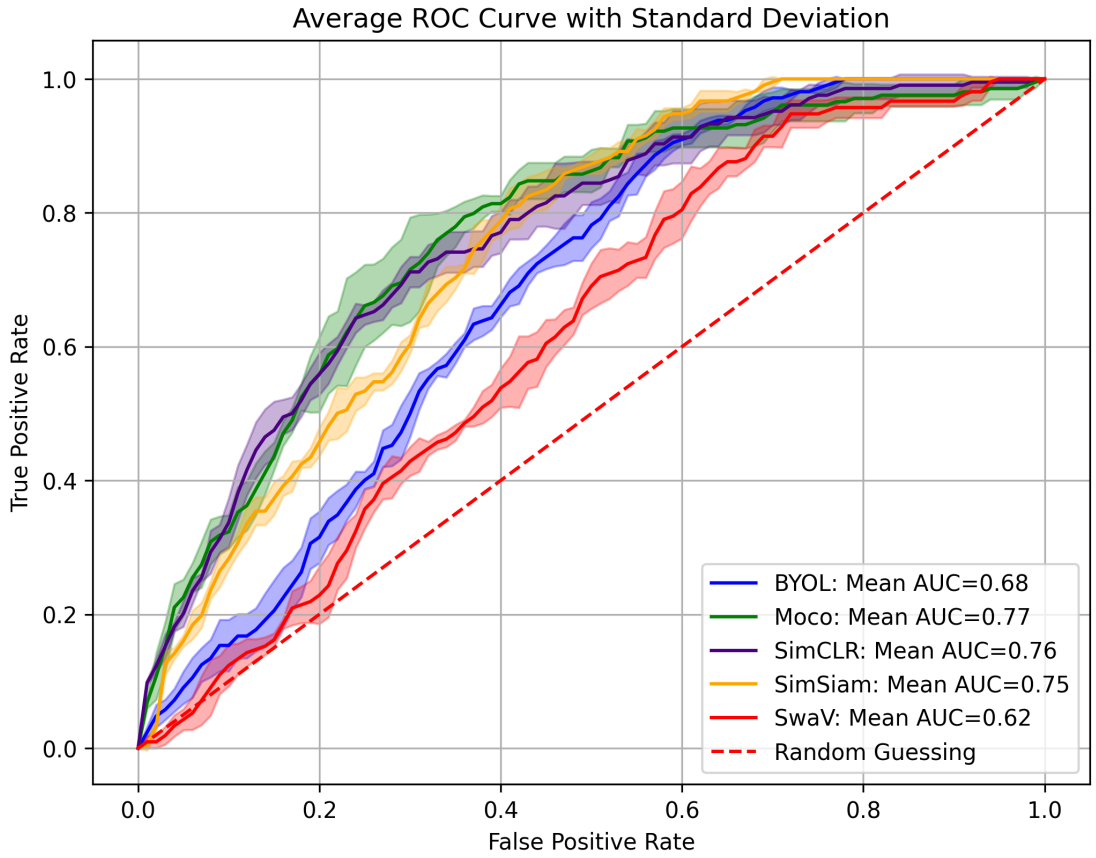

FIGURE 2.11. – **Courbe ROC avec écart-type sur la détection de mélanomes en fonction de chaque encodeur.**

Les résultats dérivés de l'utilisation des caractéristiques extraites par des modèles auto-supervisés sont résumés dans les Tableaux 2.3 et 2.4, accompagnés par les courbes ROC correspondantes illustrées dans la Figure 2.11. Parmi toutes les architectures auto-supervisées, MoCo surpasse systématiquement les autres modèles, affichant une AUC de 0.768 et une précision équilibrée de 0.708, ce qui correspond à ses performances dans d'autres tâches de la littérature. En revanche, les autres modèles montrent des performances légèrement inférieures, avec des AUC variant entre 0.62

et 0.76.

Les résultats des encodeurs auto-supervisés sont indubitablement inférieurs à ceux des modèles supervisés ; cependant, il est notable que ces architectures parviennent tout de même à obtenir une précision décente malgré la génération d'embeddings sans accès à aucune annotation. A titre de comparaison avec notre premier modèle décrit dans la partie précédente, notre nouvelle approche, usant des annotations permet effectivement une amélioration des résultats, avec une AUC pour SimSiam de 0,75, là où notre précédent modèle complètement non-supervisé atteignait 0,72.

Nous pouvons remarquer de ces résultats d'une part la robustesse et l'efficacité des caractéristiques de MoCo, mais également les performances également convenables des autres approches auto-supervisées.

2.4.4. Discussion

Nous avons présenté une nouvelle architecture pour la détection du mélanome qui exploite la puissance des Set-Transformers associés à des connexions résiduelles et à des encodeurs supervisés et auto-supervisés. Notre approche s'est avérée efficace pour calculer efficacement des caractéristiques robustes et identifier les mélanomes, ce qui a entraîné une augmentation notable de l'exactitude du modèle de classification. Ces conclusions présentent une importance significative dans le domaine de la dermatologie et du diagnostic assisté par ordinateur, ouvrant la voie à de nouvelles recherches sur le VPC. Associée aux récents progrès dans les techniques de dépistage et d'imagerie corporelle complète, la recherche sur la comparaison d'images est susceptible d'avoir un impact positif significatif dans les environnements cliniques, fournissant aux dermatologues des informations supplémentaires lors du diagnostic des lésions cutanées.

Les caractéristiques acquises par le biais de modèles auto-supervisés ont le potentiel d'améliorer significativement l'interprétabilité des prédictions et de permettre une visualisation approfondie des espaces latents, éliminant ainsi le besoin de la tâche de classification d'origine. Ces capacités peuvent renforcer les dermatologues dans leur pratique clinique, notamment en ce qui concerne les techniques de visualisation qui pourraient être appliquées dans les environnements cliniques.

De plus, notre approche présente une large applicabilité au-delà de la dermatologie, s'étendant à une large gamme de maladies et de scénarios nécessitant la comparaison de plusieurs instances et leur prédiction. Sa capacité d'adaptation dans les cas où l'obtention d'annotations est difficile ou sujette au bruit en fait un outil polyvalent et puissant dans les domaines médicaux et de la santé en général. Nous pensons que ces modèles pourront être utilisés dans le diagnostic de nombreuses autres maladies.

2.5. Synthèse

La comparaison des lésions a permis d'améliorer la qualité du diagnostic du mélanome. Nous avons examiné diverses approches actuelles qui sont capables au-

jourd'hui de comparer efficacement les lésions cutanées, ainsi que les approches permettant un diagnostic fiable pour la détection du mélanome. Nous avons pu voir que des modèles entièrement non-supervisés étaient effectivement capables de détecter les mélanomes uniquement de par leurs différences avec d'autres lésions. Nous avons ensuite observé que la comparaison supervisée des lésions permettait également d'améliorer les résultats en intégrant ces informations aux ConvNets. Toutefois, bien que nous ayons utilisé divers modèles d'encodage jusqu'à présent, nous ne les avons pas décrits en détail, alors que ces derniers sont une part très importante de notre recherche. Ces encodeurs, initialement conçus comme de simples CNN de classification, deviennent en réalité, au fil des ans, des modèles uniques en leur genre, désormais appelés Modèles d'Apprentissage de Représentations Visuelles. Ce domaine de recherche reste encore très ouvert, car les modèles existants de CNN ont un biais inductif par leur architecture, ce qui peut contraindre l'apprentissage de ces modèles de vision. De plus, dans le cas du mélanome, les CNN de classification cherchent à discriminer efficacement les mélanomes des lésions saines ; cependant, le fonctionnement interne des CNN reste obscur et peut se reposer sur des biais de classification ne reflétant pas nécessairement les images d'une manière utile pour les humains, mais plutôt d'une manière qui permet au modèle de générer des représentations facilitant la classification. La conclusion des travaux de cette partie nous a amenés à remettre en question les capacités d'encodage de nos modèles de classification, non seulement dans le cadre du mélanome mais également dans celui de la vision par ordinateur en général.

Ainsi, dans le chapitre suivant, nous explorerons les pistes que nous avons choisies pour permettre une manière différente d'encoder les images. L'idée générale est de permettre une représentation des images sous forme vectorielle qui puisse générer des représentations différentes de celles qu'un modèle de classification pourrait initialement produire. Nous verrons comment certaines approches reposent sur des modèles dont le mode de fonctionnement diffère radicalement des modèles de classification, mais qui permettent néanmoins des représentations diverses et de qualité, le tout de manière essentiellement non-supervisée ou auto-supervisée.

3. Apprentissage auto-supervisé de représentations visuelles

Sommaire

3.1. Introduction . . . 72
3.2. Etat de l'art . . . 74
3.2.1. Origines du domaine . . . 74
3.2.2. Méthodes actuelles . . . 75
3.2.3. Limites et perspectives . . . 80
3.3. Conception d'un nouveau modèle d'apprentissage de représentations visuelles . . . 81
3.3.1. Architecture proposée . . . 81
3.3.1.1. Architecture générale. . . . 82
3.3.1.2. Tâche de reconstruction. . . . 82
3.3.1.3. Tâche de Puzzle. . . . 83
3.3.1.4. Tâche Siamoise. . . . 83
3.3.2. Expériences sur Caltech-256 . . . 84
3.3.2.1. Base de données . . . 84
3.3.2.2. Méthode Expérimentale . . . 85
3.3.2.3. Équilibrage des fonctions de perte . . . 85
3.3.2.4. Études ablatives . . . 86
3.3.2.5. Impact du ratio de masquage . . . 87
3.3.2.6. Images reconstruites . . . 89
3.3.3. Expériences sur les lésions cutanées . . . 90
3.3.3.1. Résultats . . . 90
3.3.3.2. Images reconstruites . . . 91
3.3.4. Discussion . . . 91
3.4. Synthèse . . . 92

3.1. Introduction

Dans les chapitres précédents, nous avons exploré plusieurs méthodes existantes pour extraire des vecteurs d'images et les avons comparées en termes de performances sur nos bases de données. Bien que cette comparaison ait été instructive, nous ne nous sommes pas intéressés aux fonctionnements internes de ces modèles ni aux différences fondamentales qui les distinguent.

Après une analyse approfondie des différentes techniques de comparaison d'images, nous avons souhaité nous concentrer sur un aspect plus théorique et fondamental de notre approche : les modèles d'apprentissage de représentations visuelles. Auparavant, nous avons utilisé des architectures telles que SimSiam ou simplement des CNN de classification pour extraire des vecteurs des images. Bien que ces approches soient pratiques, elles présentent des inconvénients notables : L'utilisation de CNN de classification induit un biais en faveur de la tâche principale de classification, alors que la comparaison d'images devrait idéalement se baser sur des caractéristiques fondamentales de l'image. De plus, les modèles dont l'apprentissage diffère de la classification classique peuvent offrir une diversité d'analyses des images, permettant une comparaison plus poussée et complexe.

Dans ce chapitre, nous examinerons en détail les différentes méthodes existantes et la manière dont le domaine de l'apprentissage de représentations visuelles s'est développé au cours des dernières années. Nous présenterons également nos travaux qui ont abouti à la conception d'un nouveau modèle innovant utilisant des patchs des images, que nous utiliserons pour le diagnostic du mélanome. Contrairement aux réseaux neuronaux de classification, qui excellent dans une tâche spécifique, les modèles auto-supervisés visent à générer des plongements robustes et adaptables à une variété de tâches ultérieures. Les plongements ainsi obtenus sont voués à être utilisable pour de nombreuses tâches subséquentes, qui peuvent aussi bien comprendre la classification, que la comparaison d'instances, la détection d'anomalies et bien d'autres problèmes. Cette adaptabilité est cruciale dans une ère où la capacité à traiter et interpréter de vastes ensembles de données diversifiés est primordiale.

Plus particulièrement, les méthodes auto-supervisées connaissent un engouement croissant car elles permettent l'apprentissage à partir de données non étiquetées, contournant ainsi les défis liés à l'acquisition d'annotations. Les modèles auto-supervisés existants ont largement utilisé des réseaux siamois et l'augmentation de données pour apprendre à partir de données non étiquetées, posant les bases des modèles sans supervision explicite. Nous verrons que les différents modèles présents dans la littérature arborent en réalité une diversité de mécanismes, et il est certain que de nouvelles méthodes fleuriront au fil des années. Ce domaine semble donc riche en opportunités d'innovations.

Inspirés par ces travaux passés, nous proposons une architecture nouvelle issue de ces méthodologies. Nous présentons un nouveau modèle d'apprentissage de représentations visuelles, Siamese Jigsaw Masked AutoEncoder (SJ-MAE). Notre modèle combine différentes tâches afin de mieux représenter les images dans l'espace latent. Les tâches en question sont basées sur l'usage de patchs des images, à savoir la reconstruction d'image, la résolution de puzzles et les tâches d'apprentissage siamois. Ces trois tâches ont récemment obtenu des résultats prometteurs que nous analyserons dans la revue de la littérature. Notre modèle vise ainsi à hériter des forces de chacune de ces approches individuelles, les amplifiant par leur combinaison, permettant ainsi d'apprendre des plongements plus complets et robustes. Cette approche promet non seulement d'égaler mais de surpasser les modèles existants en termes de génération de plongements mieux adaptés à une large gamme d'applications, dont la détection

du mélanome.

Enfin, bien que théorique, cette approche est adaptable à une grande diversité de problèmes et de bases de données variées. Nous présenterons les résultats obtenus d'abord sur la base de données Caltech-256, contenant tout type d'images, puis sur la base de données d'images dermatologiques ISIC 2020.

3.2. Etat de l'art

3.2.1. Origines du domaine

Le domaine de l'apprentissage de représentations visuelles se réfère à la capacité des réseaux de neurones à traiter et comprendre les informations présentes dans les images. L'idée principale est que si un réseau de neurones est suffisamment entraîné pour comprendre les images, il pourra produire des sorties efficaces pour réaliser d'autres tâches subséquentes, telles que la classification, la comparaison d'images, la réidentification de personnes ou d'objets, entre autres.

Les premières approches d'apprentissage de représentations visuelles, telles que les CNN classiques, n'étaient pas initialement conçues avec cet objectif principal en tête. À l'origine, l'importance de ces modèles résidait uniquement dans leurs sorties, à savoir les prédictions de la tâche de classification. Néanmoins, l'information contenue dans les couches intermédiaires des réseaux a très rapidement également été analysée et utilisée en pratique. En effet, les réseaux convolutifs, une fois entraînés sur des bases de données diverses, contiennent inévitablement dans leurs premiers filtres la capacité de détecter les formes simples. Les couches suivantes utilisent alors ces informations pour s'affiner et permettre ainsi la classification d'objets à des niveaux égalant ceux des humains. Les modèles entraînés sur ImageNet, par exemple, qui est l'une des plus grandes et diverses bases de données d'images, ont alors très rapidement été utilisés pour initialiser d'autres modèles afin de grandement accélérer leur entraînement, même sur la détection d'objets n'ayant rien à voir avec ce que les poids initiaux détectaient (dans une certaine mesure, bien sûr).

Ces premiers développements ont amené la recherche vers la création de modèles dont l'unique but est d'acquérir des poids adaptables afin d'initialiser d'autres modèles et de permettre un affinage (ou fine-tuning) sur tous types de problèmes. Dans la littérature, ces derniers sont référencés sous le terme de **modèles fondamentaux**, et dans le cas de la vision, plus particulièrement, de **modèles de représentations visuelles**. Dans notre cas, pour la classification et la comparaison des lésions cutanées, ces modèles pré-entraînés peuvent ensuite permettre un affinage rapide des poids et, en quelques époques, offrir un modèle performant et prêt à l'emploi.

Nous verrons par la suite la multitude d'approches existantes dans ce domaine, car l'apprentissage de représentations visuelles efficaces et robustes n'est en réalité pas un problème trivial. Représenter les images dans un espace réduit semble très efficace pour transférer les informations apprises, mais comment faire apprendre à un modèle à générer ces informations ? Il n'existe pas de méthode optimale pour le moment, mais

l'approche actuelle communément utilisée consiste à créer des "tâches prétextes" dont le but est de forcer le modèle à comprendre les images afin de les représenter correctement dans l'espace latent. Ces tâches sont par exemple la reconstruction d'images dont des parties sont volontairement masquées, la comparaison d'images entre plusieurs modèles, ou encore la résolution de puzzles. Toute la difficulté réside dans le fait que le modèle cherche toujours à résoudre les problèmes donnés de manière "naïve", et il est alors primordial de créer des tâches dont la difficulté est suffisante pour que le modèle ne puisse pas trouver de solutions faciles, à savoir sans correctement prendre en compte ce que représentent les images.

L'un des avantages de ces approches est qu'il est possible, voire même préférable, de faire des entraînements non supervisés. En effet, en évitant l'utilisation d'annotations, il devient envisageable de concevoir des modèles qui se préentraînent sur un maximum d'images récupérées sur Internet, sans nécessité de disposer d'annotations précises, souvent difficiles à obtenir. Ces méthodes jouent également un rôle crucial dans les domaines de la comparaison d'images et de la réidentification de personnes.

Bien que l'interprétabilité ne soit pas la question principale de ces modèles, leur espace vectoriel pourrait potentiellement être visualisé et analysé. Toutefois, cette approche reste encore peu présente dans la littérature et est difficilement exploitable.

3.2.2. Méthodes actuelles

Ce nouveau domaine a connu une popularité croissante ces dernières années, notamment grâce au développement de nombreuses tâches prétextes permettant d'apprendre des représentations visuelles. Ces tâches varient grandement, allant de la reconstruction de patchs d'images [79], développée en 2015, à la colorisation d'images en 2016 [80], et à la prédiction de rotations en 2018 [81]. Lors de l'entraînement, le modèle se concentre uniquement sur la résolution de tâches prétextes, telles que la reconstruction de patchs d'images qui ont été masqués ou la correction de l'orientation d'images préalablement tournées incorrectement. Il a été démontré que les modèles formés exclusivement sur ces tâches, une fois affinés pour apprendre des tâches de classification, permettent au modèle de converger beaucoup plus rapidement.

C'est également en 2018 que le domaine a regagné en popularité, avec la publication de **Contrastive Predictive Coding (CPC)** [82], qui a été l'un des précurseurs dans cette série d'innovations. CPC a la particularité de pouvoir fonctionner sur tout type de modalité comme les audios, images et vidéos. Dans le cas des images, son fonctionnement consiste à encoder différentes parties de l'image dans un certain ordre, puis à prédire l'emplacement, dans l'espace vectoriel, de la suite des autres parties de l'image, sans les voir. Pour ce faire, le modèle utilise une approche par contraste en comparant les deux vecteurs. Ce modèle a mis en avant l'importance créer des modèles adaptables en utilisant des nouveaux paradigmes d'apprentissage, posant une base pour les développements ultérieurs. Son approche non-supervisée (ou plus précisément, auto-supervisée) basée sur la prédiction même de l'image, a été très utilisée pour les années suivante dans la littérature.

Momentum Contrast (MoCo) [78], développé en 2019, a ensuite introduit un méca-

nisme novateur basé sur la conservation d'une mémoire des représentations passées, utilisée pour stabiliser l'apprentissage en continu sans nécessiter de grands lots de données. Ce modèle utilise un encodeur à momentum pour maintenir la cohérence des représentations sur des batchs étendus, permettant un apprentissage plus stable et robuste des caractéristiques visuelles. Encore une fois, il s'agit d'une approche par contraste, mais dont l'usage du momentum permet de rendre le modèle plus stable. Une limite cependant significative de ce modèle est qu'il requiert de très grands batchs lors de l'entraînement, nécessitant une puissance computationnelle très importante qui n'est pas accessible à tous. La Figure 3.1 représente le schéma de leur architecture.

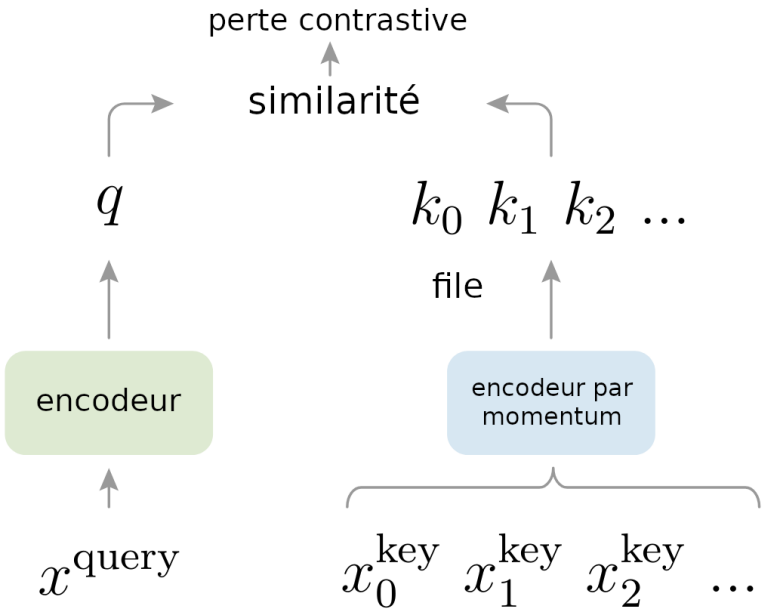

FIGURE 3.1. – **Schéma représentant l'architecture MoCo.** L'encodeur principal extrait des caractéristiques des images d'entrée et est mis à jour à chaque itération du processus d'entraînement, tandis que le momentum encoder est une version stabilisée de l'encodeur principal, qui est mise à jour lentement à l'aide d'un coefficient de momentum. *Image provenant de l'article original [78].*

L'année 2020 a été très prolifique, notamment avec l'émergence de plusieurs modèles tels que SimCLR, SimSiam, BYOL et SwaV. Les modèles siamois, qui, pendant des années, n'avaient pas eu énormément de succès, se sont finalement révélés très utiles pour les tâches d'apprentissage de représentations visuelles. Leur mécanisme à deux branches facilite en effet les cas où le modèle doit apprendre seul, en se basant sur deux entrées. Parfois, les deux branches ne partagent pas les mêmes poids, mais restent toujours utiles l'une à l'autre.

SimCLR [74] a apporté une avancée significative en simplifiant et en amplifiant l'efficacité de l'apprentissage par contraste. Ce modèle tire parti de techniques d'augmentation des données sophistiquées et de la maximisation de l'information mutuelle

entre différentes vues augmentées de la même image, réalisées à travers des batches de taille considérable. SimCLR a démontré que la qualité des représentations pouvait être fortement améliorée grâce à ces méthodes d'augmentation.

SimSiam [72] est le modèle que nous avons majoritairement utilisé pour nos premiers travaux du chapitre précédent et que nous avons expliqué précédemment.

Enfin, la même année, Bootstrap Your Own Latent (BYOL) [75] et SwAV [73] ont exploré des approches alternatives à l'apprentissage par contraste sans exemples négatifs. BYOL, en particulier, a innové en utilisant deux réseaux qui interagissent pour prédire les caractéristiques de l'un à partir de l'autre, sans nécessiter la comparaison avec des exemples négatifs, évitant ainsi les problèmes liés à l'échantillonnage de ces paires. SwAV, quant à lui, a utilisé un clustering des représentations pour réaliser un échange de prototypes entre différentes vues, une méthode qui permet une meilleure généralisation et une meilleure compréhension des données. La Figure 3.2 présente l'architecture Byol.

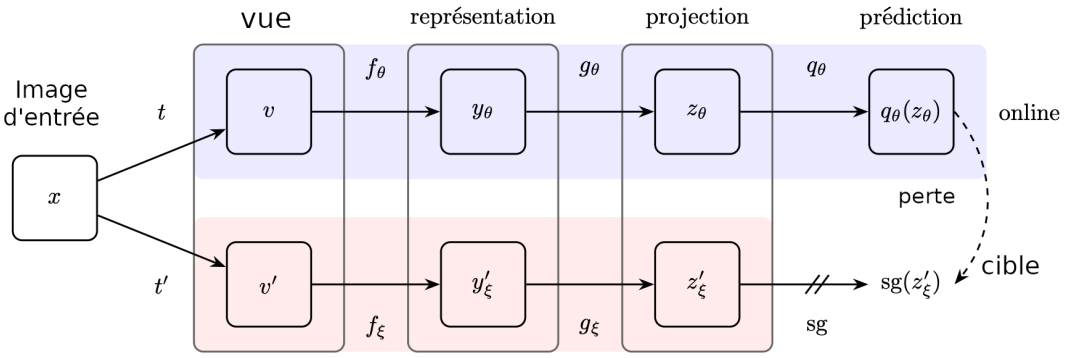

FIGURE 3.2. – **Schéma représentant l'architecture Byol.** Chacune des branches reçoit une vue augmentée d'une même image d'entrée. Le premier modèle, appelé "target network", est mis à jour de manière lente et accumulative, ce qui lui permet d'agir comme une référence stable. Le second modèle, ou "online network", est mis à jour de manière plus rapide à chaque itération.

A la même période, plusieurs innovations avaient, elles, déjà eu lieu mais cette fois dans le domaine du Traitement Automatique des Langues (TAL). En particulier, la stratégie de masquage en TAL est devenue une pierre angulaire dans le domaine, suite à sa mise en œuvre réussie dans des modèles tels que BERT (Bidirectional Encoder Representations from Transformers) [83] en 2018. Cette approche consiste à masquer aléatoirement une partie des données d'entrée — typiquement des mots ou des sous-mots — et ensuite à entraîner le modèle à prédire les éléments masqués en se basant sur leur contexte. L'efficacité de cette stratégie réside dans sa capacité à forcer le modèle à développer une compréhension profonde du contexte linguistique et des relations, apprenant ainsi des représentations robustes et polyvalentes [84]. Cette méthode d'apprentissage auto-supervisé non seulement améliore la capacité

du modèle à comprendre le langage, mais lui permet également de bien performer sur une variété de tâches subséquentes à l'aide d'un simple finetuning rapide [85].

Ainsi, les AutoEncodeurs Masqués (MAE) [86] sont apparus comme la suite logique dans le domaine de l'apprentissage auto-supervisé pour la vision par ordinateur, en s'inspirant des stratégies de masquage en TAL. Introduits en 2021 par He et al., ces derniers fonctionnent en masquant une partie significative de l'image d'entrée, typiquement environ 75 %, et en entraînant le réseau à reconstruire l'image originale à partir des parties visibles (des exemples de reconstructions sont présentés à la Figure 3.3). Bien que l'idée de masquer des parties d'une image pour la reconstruire ne soit pas nouvelle [87, 88], l'émergence des Vision Transformers (ViT) [89] a permis une manière beaucoup plus efficace de réaliser la reconstruction de parties masquées des images. En travaillant sur des patchs des images, il suffit d'ignorer les patchs masqués dans l'architecture du modèle, pour ensuite lui demander de les reconstruire, c'est l'approche utilisée dans MAE, ayant permis de très bonnes performances.

Il est important toutefois de noter que les images reconstruites par ces modèles présentent un flou induit par la fonction de perte Mean Squared Error, mais qui n'est pas problématique étant donné que le but du modèle est d'apprendre de bonne représentations visuelles, ce qui n'est pas toujours lié à des reconstructions réalistes. Cette méthode de reconstruction tire parti de la redondance dans les données visuelles, obligeant le modèle à inférer le contenu manquant en se basant sur les parties non masquées et ainsi à apprendre une représentation riche et contextuelle de l'entrée. Des recherches ultérieures ont démontré l'efficacité des MAE dans diverses tâches en aval, prouvant leur capacité à générer des caractéristiques hautement généralisables avec une supervision minimale [90-92]. Inspirés par ces résultats prometteurs, nous avons utilisé en partie cette approche dans l'architecture que nous présentons dans la partie suivante.

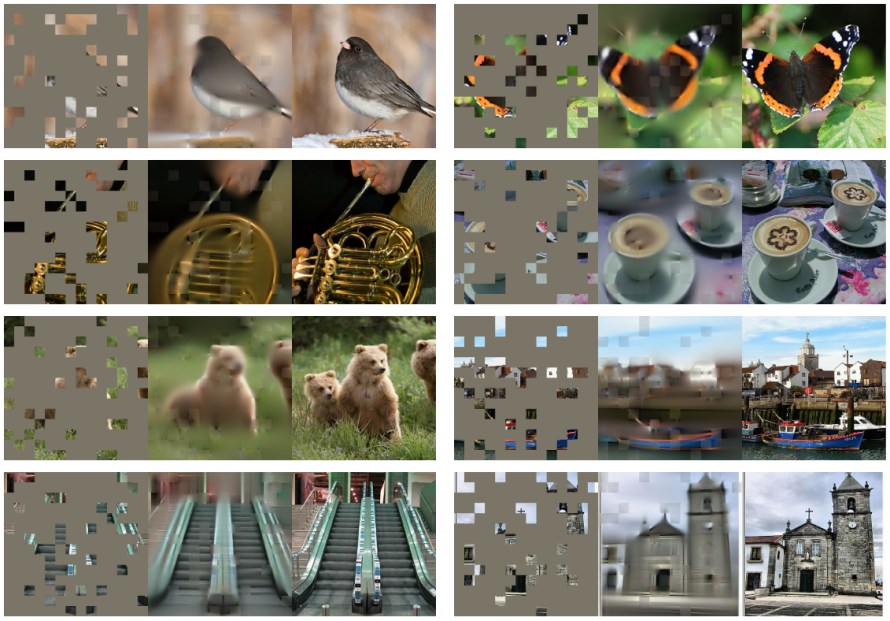

FIGURE 3.3. – **Exemple d'images reconstruites par le modèle MAE.** On constate que le modèle reste capable de reconstruire une bonne partie de l'image malgré une grande proportion de patches masqués. *Image provenant de l'article initial [86]*

Mais cette approche n'est pas la seule qui paraît prometteuse dans le domaine. En effet, les Transformers de Vision Jigsaw [93], publiés en 2023, sont apparus comme une autre manière innovante d'apprendre les représentations visuelles sans l'utilisation de patchs masqués. L'idée principale repose sur le fait que, pour résoudre des puzzles, il est nécessaire de comprendre la logique qui lie les différents patchs d'une image. En forçant le modèle à optimiser cette tâche, il est fort probable que ce dernier puisse acquérir des informations sur les images lui permettant d'être efficace sur les tâches subséquentes, sans avoir eu besoin d'annotations. Le papier initial utilise cette stratégie comme moyen d'équilibrer la généralisation et la robustesse d'un modèle de classification, plutôt que d'utiliser purement le puzzle comme méthode d'apprentissage de représentation. La perte optimisée est donc la somme de l'entropie croisée de classification et de l'entropie croisée des embeddings de localisation prédits pour la tâche du puzzle.

Enfin, il existe également les réseaux neuronaux siamois. Ces derniers ont été utilisés pendant des années pour comparer des images, mais leurs équivalents auto-supervisés ont montré leurs avantages seulement récemment lorsqu'ils sont appliqués à la tâche de pré-entraînement. Ils reposent sur la comparaison de multiples instances, souvent calculée comme la distance entre ces instances. En apprentissage supervisé, les étiquettes indiquent si deux images sont similaires ou non, et le modèle apprend à minimiser ou maximiser la distance entre les deux embeddings obtenus via un CNN de base. La perte de similarité utilisée est souvent la similarité cosinus (négative dans le cas d'échantillons négatifs), comme le montre l'équation 3.1.

$$SimilaritéCosinus(\mathbf{z_1}, \mathbf{z_2}) = \frac{\mathbf{z_1} \cdot \mathbf{z_2}}{||\mathbf{z_1}|| \cdot ||\mathbf{z_2}||} \tag{3.1}$$

Dans le contexte de l'apprentissage auto-supervisé, de nombreux travaux ont montré que les approches siamoises étaient également capables d'apprendre une représentation visuelle utile, même en utilisant très peu d'opérations [72].

Actuellement, les modèles siamois semblent être légèrement éclipsés par des techniques plus populaires, mais ils restent néanmoins une méthode efficace pouvant être exploitée dans l'apprentissage de la représentation visuelle. En concevant une architecture exclusivement pour les vidéos, Gupta et al. [94] ont montré que leur architecture siamoise (SiamMAE) surpassait un large panel de méthodes auto-supervisées de pointe sur les tâches de segmentation d'objets vidéo, de propagation de points clés de pose et de propagation de parties sémantiques. D'autres travaux axés sur les images ont également permis d'obtenir de très bonnes performances en fusionnant une configuration contrastive avec la tâche de reconstruction [95, 96].

3.2.3. Limites et perspectives

Ces nouvelles méthodes présentent une impressionnante diversité de mécanismes différents pour apprendre les représentations visuelles. Parmi toutes ces méthodes, aucune ne s'est pour l'instant démarquée comme étant supérieure aux autres, et il semble qu'il reste encore énormément de place dans ce domaine de recherche pour trouver de nouvelles solutions innovantes.

Certaines approches se reposent uniquement sur des techniques d'augmentation de données, tandis que d'autres se basent uniquement sur la tâche de reconstruction ou de résolution de puzzle. Mais est-ce que ces tâches, indépendantes, ne pourraient-elles pas travailler de pair afin d'améliorer l'apprentissage de nos modèles ? Chaque approche contient son lot de désavantages. Par exemple, les méthodes basées sur l'augmentation de données nécessitent de concevoir des augmentations efficaces tout en courant le risque de trop modifier les instances. Les approches par reconstruction sont certes performantes, mais cette unique tâche permet-elle réellement d'offrir une bonne compréhension des images ?

Ces défis soulignent le besoin de méthodes alternatives, permettant une compréhension plus riche des images. Dans cette optique, notre recherche s'oriente vers la création d'une nouvelle architecture visant à faire apprendre aux modèles la résolution de plusieurs problèmes complexes de manière simultanée. Cette optimisation multiple poussera ainsi le modèle à représenter des caractéristiques plus diverse des images.

3.3. Conception d'un nouveau modèle d'apprentissage de représentations visuelles

3.3.1. Architecture proposée

Nous développons une architecture auto-supervisée, appelée Siamese Jigsaw Masked AutoEncoders (SJ-MAE), optimisant simultanément les multiples tâches de reconstruction, de résolution d'un puzzle et de rapprochement des embeddings des patches de la même image. Ces trois pertes sont optimisées à l'aide du même encodeur, qui sera utilisé pour les tâches subséquentes (dans notre cas, la classification et comparaisons d'images). Le décodeur pour la tâche de reconstruction et le MLP pour la tâche Puzzle représentent deux branches différentes du modèle qui sont optimisées indépendamment, mais qui tirent toutes deux parti de l'encodeur commun. La branche siamoise utilise également l'encodeur principal, auquel un prédicteur est ensuite ajouté.

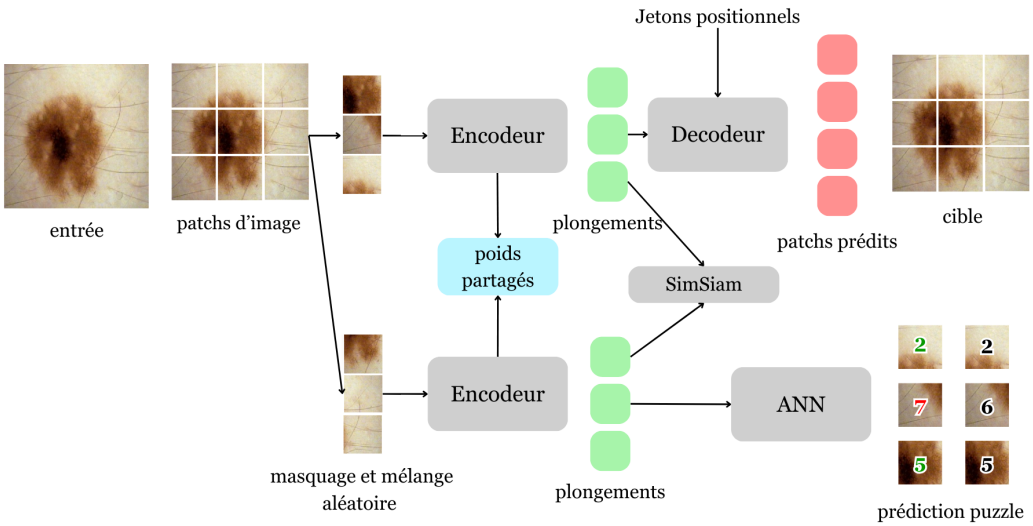

FIGURE 3.4. – **Vue d'ensemble de l'architecture SJ-MAE.** Un MAE classique est implémenté avec ses modules encodeur-décodeur. Nous ajoutons deux nouvelles branches correspondant à la tâche de puzzle et à la tâche siamoise. Le module puzzle prend les plongements extraits de l'encodeur, sans embeddings positionnels, et résout le puzzle sur les patches visibles, tandis que la branche siamoise rapproche les deux plongements l'un de l'autre.

3.3.1.1. Architecture générale.

Nous avons d'abord mis en œuvre un MAE inspiré par l'article de Chen et al. [86], qui consiste en un Transformer Visuel (ViT) avec un encodeur et un décodeur. L'encodeur prend en entrée un sous-ensemble de patches d'une image, tandis que le décodeur reconstruit l'image originale en utilisant les embeddings dérivés de l'encodeur, reconstruisant les patchs manquants. Nous avons conservé les paramètres par défaut et l'architecture principale du ViT de base. Lors de l'entraînement de cette branche, nous fournissons les embeddings positionnels au modèle, contrairement à la tâche puzzle, qui doit les prédire. Étant donné que les valeurs d'entrée diffèrent entre les deux branches, nous employons une tâche siamoise pour rapprocher les deux représentations dans l'espace latent, car elles représentent la même image, mais masquée différemment. Cela implique d'aligner les représentations de deux sous-ensembles différents de patchs de la même image — un ensemble avec des embeddings positionnels et l'autre sans. Ensuite, lors de l'affinage (ou fine-tuning) et de la classification linéaire (aussi appelée linprobe), qui seront nos deux méthodes de validation, les embeddings positionnels seront naturellement fournis au modèle.

3.3.1.2. Tâche de reconstruction.

En utilisant l'erreur quadratique moyenne (MSE), à la fois l'encodeur et le décodeur sont optimisés pour réduire la différence entre l'image d'entrée et l'image prédite. Similaire à l'implémentation originale de MAE, nous ne calculons pas la perte sur les patches visibles, mais nous nous concentrons uniquement sur les patches masqués, ce qui entraîne une qualité inférieure pour la reconstruction des patches visibles. Dans l'article original sur le MAE, une conséquence intéressante est que ces patches visibles, pour lesquels il n'y a pourtant aucun calcul de gradient, sont tout de même reconstruits en partie par le décodeur. Cela laisse à penser que le modèle génère à la fois une représentation globale de l'image ainsi qu'une représentation abstraite des patches donnés. Dans notre architecture, l'encodeur, représenté par ϕ, projette les données d'entrée $\mathbf{x}$ avec les embeddings positionnels $\mathbf{E_{pos}}$ vers une représentation latente $\mathbf{z}$. Nous conservons les valeurs des embeddings positionnels de l'article original, qui sont non optimisés et constants grâce à une fonction sin-cos. Le processus peut être formellement exprimé comme suit :

$$\mathbf{z} = \phi(\mathbf{patchify}(\mathbf{x}) + \mathbf{E_{pos}})$$

Le décodeur, représenté par ψ, projette ensuite l'embedding latent $\mathbf{z}$, concaténé avec des jetons de masque $[MSK]$, à l'image originale. Les embeddings positionnels $\mathbf{E_{pos}}$ sont également ajoutés à chaque embedding.

$$\hat{\mathbf{x}} = \psi((\mathbf{z}.[MSK]) + \mathbf{E_{pos}}) \tag{3.2}$$

La perte de reconstruction est ensuite décrite par l'équation 3.3.

$$\mathscr{L}_r = MSE(\hat{\mathbf{x}}, \mathbf{x}) \tag{3.3}$$

3.3.1.3. Tâche de Puzzle.

Dans la tâche de puzzle, un nouvel ensemble de patchs visibles de la même image est mélangé et passé à l'encodeur, sans leurs embeddings positionnels. Les embeddings issus de l'encodeur sont donnés en entrée à un petit réseau de neurones, optimisé pour minimiser l'entropie croisée entre les positions prédites et les positions cibles. Ce réseau de neurones cherche donc à prédire la position des patchs visibles en fournissant en sortie les embeddings positionnels prédits. Nous désignons les positions cibles par $\mathbf{p}$ et les positions prédites par $\hat{\mathbf{p}}$. L'architecture du réseau Puzzle est composée de trois couches entièrement connectées, désignées par H. Ainsi, la perte pour cet objectif est l'entropie croisée, notée CE, et est décrite dans l'équation 3.4.

$$\mathscr{L}_j = CE(H(\mathbf{z}), \mathbf{p}) \tag{3.4}$$

3.3.1.4. Tâche Siamoise.

La branche siamoise est inspirée par l'architecture SimSiam [72]. Nous exploitons le même encodeur que celui utilisé dans les autres branches, puis nous alimentons les sorties vers la branche siamoise. Cette branche consiste en un simple prédicteur (couche de réseau de neurones) prenant à chaque étape un seul embedding, tandis que l'autre branche est dirigée vers le calcul de la perte sans passer par le prédicteur. Une modification importante que nous avons ajoutée par rapport à l'article original est que, au lieu d'utiliser des images entières, nous fournissons au modèle deux sous-ensembles différents de patchs de l'image. Les deux branches subissent des transformations colorimétriques aléatoires faibles. Faisant écho à l'idée centrale de SimSiam, qui tire parti des techniques d'augmentation de données randomisées, le choix d'utiliser des sous-ensembles de patchs sert de stratégie d'augmentation supplémentaire. Cette approche augmente probablement la difficulté de la tâche, aidant ainsi le modèle à apprendre des embeddings plus robustes.

Cependant, nous n'avons pas employé toutes les stratégies d'augmentation de SimSiam, initialement développées pour l'article MoCo [78]. Ces transformations, combinées à la stratégie de masquage, enlèvent trop d'informations critiques des images, notamment dans le cas des lésions cutanées.

Formellement, cette branche est composée d'un prédicteur $\mathscr{P}$, qui inclut deux couches entièrement connectées et une opération de stop-gradient pour protéger le modèle contre la dégénération. La perte siamoise est définie par :

$$\mathscr{L}_s = \mathscr{D}(stopgrad(\mathbf{z_1}), \mathscr{P}(\mathbf{z_2})) \tag{3.5}$$

avec $\mathscr{D}$ la distance cosinus et $\mathbf{z_1}$, $\mathbf{z_2}$ les deux embeddings provenant des différents sous-ensembles de patchs de l'image, ayant subi différentes augmentations de données. **Ainsi, la perte totale** est donc décrite comme suit :

$$\mathscr{L}_{total} = \mathscr{L}_r + \mu\mathscr{L}_j + \alpha\mathscr{L}_s \tag{3.6}$$

avec μ et α deux hyperparamètres représentant les poids correspondants pour les pertes puzzle et siamoise.

3.3.2. Expériences sur Caltech-256

Tout d'abord, nous avons souhaité valider les performances de notre architecture dans un cas général. Ainsi, nous avons réalisé nos premières expériences sur une base de données représentant une vaste diversité d'éléments, à savoir la base de données Caltech-256. Sur cette base, nous avons effectué plusieurs études ablatives et analysé le comportement de notre modèle en fonction de différentes valeurs d'hyperparamètres. Nous présenterons dans la partie suivante nos résultats sur la détection du mélanome.

3.3.2.1. Base de données

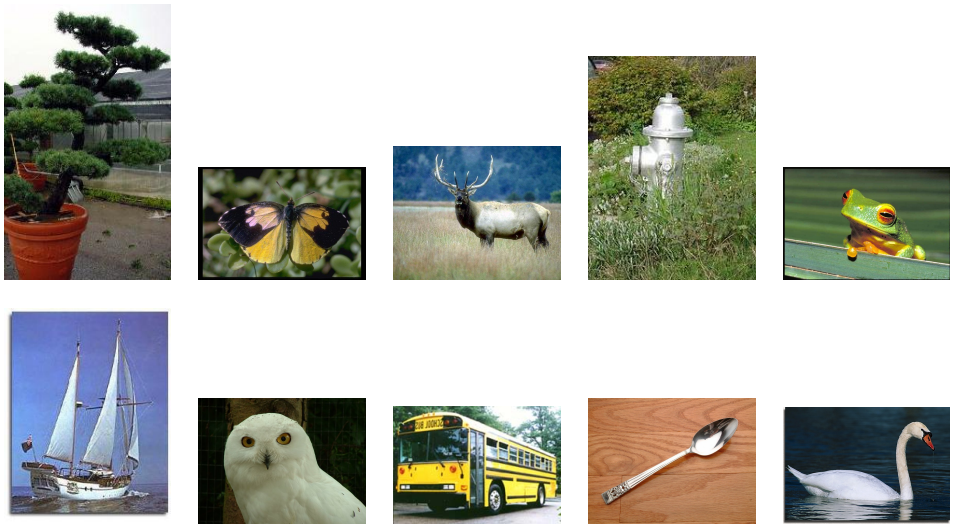

FIGURE 3.5. – **Exemple d'images de Caltech-256.**

Le dataset Caltech-256 [97] (voir les exemples Fig. 3.5) est une version étendue du précédent dataset Caltech-101, visant à fournir un benchmark plus difficile pour les modèles de vision par ordinateur. Ce dataset comprend 30 607 images réparties en 256 catégories d'objets, chacune contenant au moins 80 images, plus une catégorie supplémentaire pour les images qui ne s'intègrent pas nettement dans les autres catégories. La diversité et l'augmentation du nombre de catégories dans Caltech-256 en font un choix approprié pour tester la robustesse et l'efficacité des algorithmes d'apprentissage auto-supervisé. La diversité et le déséquilibre entre les classes, où certaines catégories contiennent nettement moins d'images que d'autres, posent des défis particuliers pour les modèles d'apprentissage automatique, notamment ceux basés sur des techniques d'apprentissage auto-supervisé. Ces conditions compliquent souvent le processus d'entraînement, rendant difficile l'obtention d'une haute précision à travers toutes les classes, mais permettant ainsi d'analyser les performances de notre modèle.

Dans nos expériences, nous explorons l'efficacité des méthodes d'entraînement auto-supervisées pour relever ces défis à travers deux méthodes de validation. Nous vérifions les performances des poids appris en entraînant un modèle linéaire d'une part, puis en affinant le modèle à la tâche de classification afin de comparer les performances. Nous avons divisé aléatoirement la base de données en un ensemble d'entraînement et de test contenant respectivement 80 % et 20 % des données.

3.3.2.2. Méthode Expérimentale

Nous adhérons à la pratique courante dans l'apprentissage auto-supervisé de représentations, en commençant initialement par un entraînement auto-supervisé comme étape de pré-entraînement. Nous utilisons ensuite les poids calculés pour entraîner un classificateur linéaire et affiner le modèle. La phase de pré-entraînement dure 800 époques, incluant 10 époques pour lesquelles le taux d'apprentissage est faible et croît lentement. Nous utilisons l'optimiseur AdamW avec un taux d'apprentissage de base de 0,002, une décroissance de poids de 0,05, et une taille de lot de 64. Pendant le finetuning, le modèle est entraîné pour 50 époques supplémentaires, et nous conservons les hyperparamètres par défaut tels que définis dans l'article original MAE [86]. La classification linéaire est effectuée sur 90 époques avec les hyperparamètres par défaut. Pour nos expériences, le modèle de base utilisé est un ViT base patch 16 [98].

3.3.2.3. Équilibrage des fonctions de perte

Si elles ne sont pas correctement équilibrées, les pertes sélectionnées peuvent présenter des magnitudes disparates, potentiellement perturbatrices pour le processus d'entraînement du modèle. Nos expériences initiales ont été conçues pour analyser le comportement des différentes pertes pendant l'entraînement. Après avoir effectué 2 000 étapes d'entraînement, nous avons examiné les valeurs moyennes de chaque perte. Comme illustré dans le Tableau 3.1, il est évident que la perte du puzzle est significativement plus grande que la perte de reconstruction — par un facteur d'environ sept — tandis que la perte contrastive reste comparativement faible. En réponse à ces découvertes, nous avons ajusté les poids pour nos sessions d'entraînement ultérieures : fixant un poids pour la perte du puzzle à 0.1 et conservant un poids de 1 pour la tâche contrastive.

Etape	Perte de reconstruction	Perte puzzle	Perte siamoise	Total
500	1.25	5.27	0.08	6.61
1,000	0.67	5.22	0.01	5.95
1,500	0.62	5.18	0.02	5.87
2,000	0.52	5.18	0.02	5.74

TABLEAU 3.1. – **Valeurs des différentes pertes en fonction du nombre d'étapes.** En moyenne, la perte du puzzle est 7 fois plus élevée que la perte de reconstruction.

Lors de l'ajustement des poids des pertes, les valeurs de perte diminuent de manière appropriée au fil des époques, comme illustré dans la Figure 3.6, et restent dans des fourchettes raisonnables, restant en dessous de 1. De manière intéressante, la perte siamoise, en revanche, est en augmentation, ce qui est probablement dû au fait que cette perte est plus petite que les deux autres, mais elle contribue à la stabilité pendant l'entraînement en maintenant les embeddings plus proches les uns des autres.

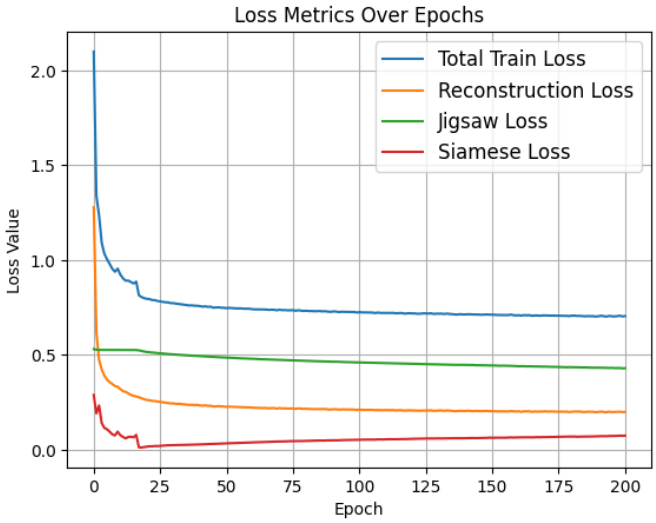

FIGURE 3.6. – **Valeurs des fonctions de pertes une fois équilibrées.**

3.3.2.4. Études ablatives

L'optimisation de plusieurs objectifs peut être un défi pour les réseaux neuronaux. Cela introduit de la complexité, peut ralentir l'apprentissage, mais peut également permettre aux modèles de développer une compréhension plus nuancée des données qu'ils traitent. Dans cette section, nous comparons les performances de différentes architectures : un Vision Transformer (ViT) classique pour la classification, une architecture d'Autoencodeur Masqué (MAE), l'ajout de la tâche Puzzle (J-MAE), et notre approche, SJ-MAE, contenant la tâche Puzzle et Siamoise. Tous les modèles masqués utilisent un modèle de base ViT avec des patches de taille 16 [98], entraînés à partir de zéro pendant 800 époques avec un ratio de masquage de 70 %. Pour la tâche de classification, nous avons opté pour un modèle de base DeiT avec des patches de taille 16 [99], qui a surpassé le ViT original et étant plus rapide. Nous avons initialement entraîné le DeiT de classification pendant 100 époques et l'avons ensuite peaufiné pour la validation linéaire dans les mêmes conditions que les modèles masqués. Les résultats sont présentés dans les Figures 3.2 et 3.7.

D'après les résultats, il est clair que les architectures J-MAE et SJ-MAE surpassent le ViT de Classification de base et le MAE standard sur tous les critères évalués. Plus précisément, J-MAE et SJ-MAE démontrent une performance supérieure non seulement dans les scénarios de fine-tuning, où ils atteignent des précisions Top-1 de 66,76 % et

Architecture	Linprobe Top-1	Linprobe Top-5	Finetune Top-1	Finetune Top-5
Classification ViT	19.96	36.10	36.36	55.95
MAE	23.13	40.62	59.34	77.61
J-MAE	**27.77**	**48.66**	66.76	82.88
SJ-MAE	27.47	48.44	**67.24**	**83.42**

TABLEAU 3.2. – **Précision des variantes MAE.** Les meilleures précision par colonne sont affichée en gras.

67,24 % respectivement, mais aussi en classification linéaire, où J-MAE enregistre une précision Top-1 de 27,77 % et SJ-MAE suit de près avec 27,47 %.

SJ-MAE montre les meilleures performances, particulièrement dans la tâche de fine-tuning, où il surpasse systématiquement toutes les autres méthodes à travers toutes les époques. Cependant, il est à noter que dans les époques ultérieures, pendant la tâche de classification linéaire, l'écart de performance entre SJ-MAE et J-MAE se réduit. À ce stade d'apprentissage, les représentations dérivées des pertes de puzzle et de reconstruction peuvent être suffisamment robustes pour se comporter de manière équivalente, avec ou sans tâches siamoises. Parmi toutes les époques, ces derniers montrent des précisions Top-5 significativement plus élevées comparées à MAE et au ViT de classification dans les contextes de classification linéaire et de fine-tuning.

L'architecture MAE elle-même montre une amélioration notable par rapport au ViT de classification standard, particulièrement dans la phase de fine-tuning, où elle atteint une précision Top-1 de 59,34 % comparée à 36,36 % pour le ViT. Cela souligne l'efficacité du MAE pour apprendre des caractéristiques généralisées qui peuvent être peaufinées pour des tâches spécifiques plus efficacement que celles apprises par le ViT de classification entraîné directement.

3.3.2.5. Impact du ratio de masquage

Les résultats illustrés dans la Figure 3.8 offrent une analyse détaillée des différences de performance en termes de précisions Top-1 et Top-5 pour les modèles utilisant des réseaux siamois par rapport à ceux qui n'en utilisent pas, à travers une gamme de ratios de masquage, dans les scénarios de fine-tuning et de classification linéaire.

Le ratio de masquage, l'un des hyperparamètres les plus importants, a été varié sur neuf niveaux différents, allant de 10 % à 90 %. Bien que des ratios de masquage plus élevés offrent des avantages tels qu'une réduction du calcul en raison de moins de données d'entrée, masquer plus de 80 % des images laisse clairement le modèle avec une quantité insuffisante d'information pour une compréhension efficace de l'image, conduisant à une chute significative des performances. Les ratios de masquage optimaux sont identifiés comme étant entre 50 % et 60 % pour la tâche de fine-tuning et entre 70 % et 80 % pour la tâche de classification linéaire.

Dans les mesures de précision Top-1 et Top-5, les modèles siamois surpassent systématiquement leurs homologues non siamois sur presque tous les ratios de masquage. Cette performance indique que les réseaux siamois sont utiles pour apprendre des

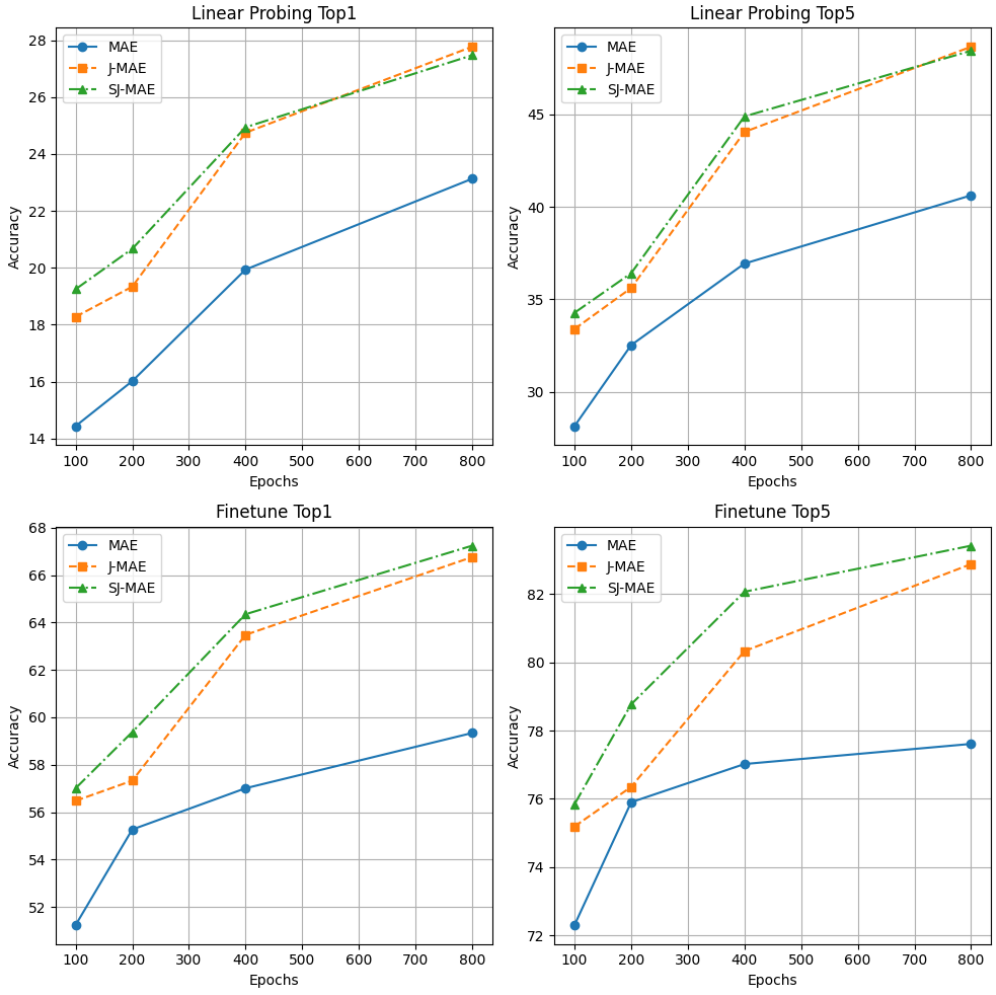

FIGURE 3.7. – **Précision des différents modèles par nombre d'époques de pré-entraînement.**

caractéristiques robustes qui améliorent les capacités de classification.

Les performances pour les deux métriques de précision culminent généralement à un ratio de masquage d'environ 0,5 dans les scénarios de fine-tuning, tandis que dans la classification linéaire, la performance maximale est légèrement plus élevée, autour de 0,8.

Au-delà de ces ratios de masquage, il y a une baisse notable des performances pour les deux métriques et scénarios. Cela suggère que des niveaux plus élevés de masquage des entrées peuvent ajouter une complexité excessive au processus d'apprentissage, tandis que des masquages d'entrée moindres peuvent rendre les tâches trop simples. De plus, il est possible que les 100 époques n'aient pas été suffisantes pour que les trois tâches soient minimisées efficacement. Cependant, cette baisse est moins prononcée dans les modèles siamois, en particulier dans la classification linéaire pour la précision Top-5, ce qui démontre leur robustesse relative et leur capacité améliorée à gérer des entrées de données ambiguës ou incomplètes.

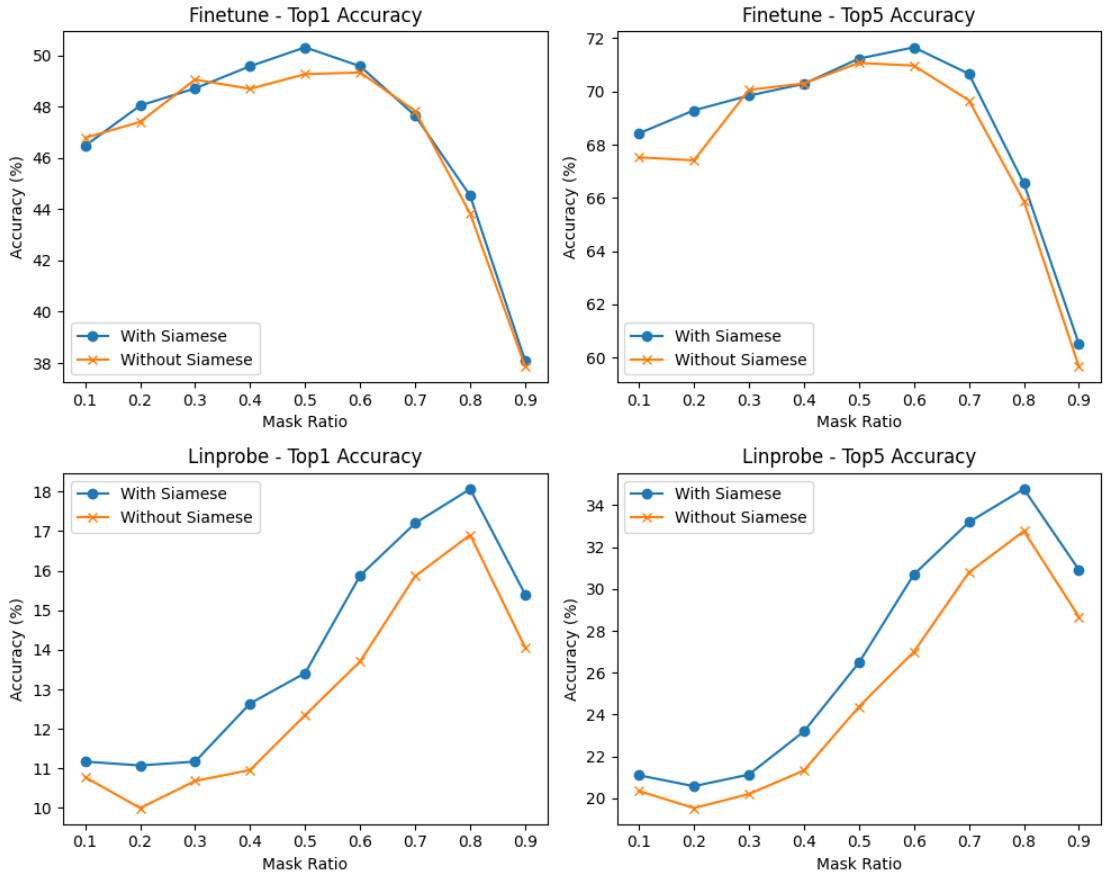

FIGURE 3.8. – **Comparaison des résultats de la classification linéaire et du fine-tuning en fonction de différents taux de masquage.** Résultats obtenus après 100 époques de pré-entraînement, suivies de 50 époques de fine-tuning ou de 100 époques de classification linéaire.

En moyenne, dans ces expériences, la tâche siamoise a conduit à une amélioration de la précision d'environ 1,2 %.

3.3.2.6. Images reconstruites

Bien que la tâche de reconstruction soit utilisée uniquement comme une tâche prétexte, le modèle final est encore capable de reconstruire efficacement des images à partir d'une petite quantité de patchs visibles. En raison de l'utilisation de la perte MSE et de l'architecture entièrement basée sur l'encodage-décodage, le modèle n'a pas été contraint de générer des images avec des détails précis, mais nous pouvons clairement voir dans les images générées que le modèle a fait des suppositions intéressantes sur les images représentées à partir des patchs. Nous présentons plusieurs exemples dans la Figure 3.9.

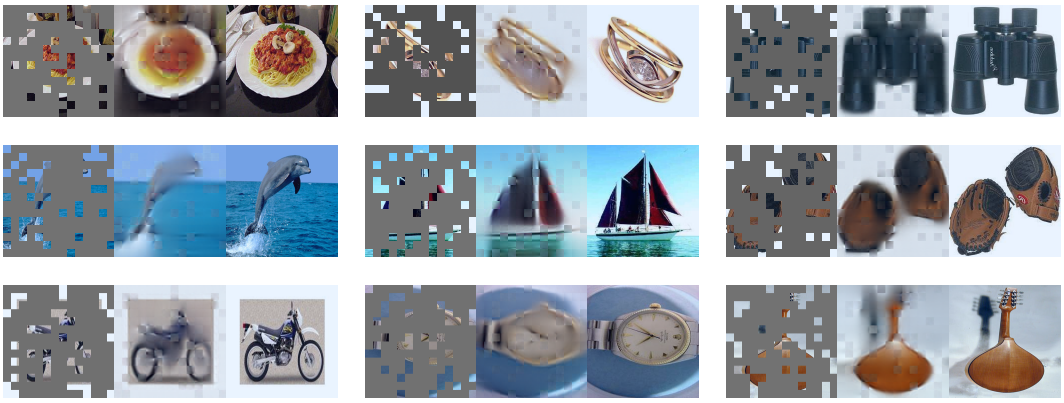

FIGURE 3.9. – **Exemples de résultats sur les images de validation de Caltech-256.** (à gauche) l'image masquée, (au milieu) la reconstruction par SJ-MAE, (à droite) vérité terrain. Le ratio de masquage est de 80 %, ne laissant que 39 patchs sur 196.

3.3.3. Expériences sur les lésions cutanées

En utilisant la même méthode expérimentale, nous pouvons analyser les résultats sur la base de données sur le mélanome, à savoir ISIC 2020. Dans le cas du mélanome, nous avons choisi un taux de masquage relativement faible en raison de la complexité de reconstruire les images dermoscopiques après l'opération de masquage. En effet, contrairement aux objets présents dans la base de données Caltech-256, les images de lésions cutanées représentent généralement des lésions uniques, souvent petites, qui risqueraient de disparaître lors du masquage. De plus, la tâche du puzzle est également plus compliquée pour les images dermoscopiques à cause de la grande quantité de peau dans les images. Ces patchs de peau sont difficiles, voire impossibles à réorganiser correctement en raison de leur grande similarité.

3.3.3.1. Résultats

Comme pour Caltech, nous avons d'abord entraîné les modèles de représentation visuelle puis utilisé leurs poids pour initialiser les modèles de classification. La classification linéaire subséquente a duré 100 époques et le fine-tuning 50 époques. Les résultats obtenus par classification linéaire et finetuning sont présentés dans la Figure 3.3.

De la même manière qu'avec la base de données Caltech, nous remarquons encore une fois que les modèles optimisant plusieurs pertes en même temps ont pu donner de meilleurs résultats lors de la tâche subséquente de classification.

Afin de profiter au maximum de la qualité des représentations apprises, nous avons également analysé leurs performances lorsqu'elles sont utilisées avec l'architecture du chapitre précédent, à savoir ReSeT. Nous montrons de nouveau les résultats que nous avions obtenus auparavant, cette fois-ci avec les nouveaux résultats obtenus à l'aide de SJ-MAE dans la Table 3.4.

Architecture	Finetune AUC
Classification ViT	0.79
MAE	0.83
J-MAE	0.85
SJ-MAE	**0.86**

TABLEAU 3.3. – **Précision des variantes MAE.** Les meilleures précision par colonne sont affichée en gras.

Architecture auto-supervisée	AUC	Précision équilibrée
BYOL	0.681	0.63
MoCo	**0.768**	0.708
SimCLR	0.764	0.683
SimSiam	0.748	0.702
SwaV	0.617	0.569
SJ-MAE	0.767	**0.712**

TABLEAU 3.4. – **Résultats de la tâche de classification sur le jeu de test utilisant les architectures auto-supervisées pour l'entraînement du CNN.**

Avec une AUC à 0,767, SJ-MAE obtient une AUC similaire à celle de MoCo, et supérieure aux autres architectures. La précision équilibrée obtenue est la plus élevée de toutes les autres architectures choisies, soulignant la capacité diagnostique obtenue lors de l'usage de SJ-MAE et ReSeT.

3.3.3.2. Images reconstruites

Nous avons également analysé la capacité du modèle à reconstruire les images dans le cadre d'ISIC 2020. Nous avons remarqué que le modèle réussissait également à reconstruire les patch manquant dans le cas des lésions cutanées. Plusieurs exemples sont présentés dans la Figure 3.10.

3.3.4. Discussion

Nous avons présenté SJ-MAE, notre modèle tirant parti de plusieurs avancées récentes dans le domaine de l'apprentissage auto-supervisé des représentations. Appliqué à Caltech-256 et à ISIC 2020, nos expériences démontrent que SJ-MAE améliore les performances lors du fine-tuning pour les tâches de classification et permet également de bons résultats lors de la comparaison des lésions avec ReSeT. Ces résultats mettent en évidence les gains potentiels en termes de performance et de qualité des embeddings lors de l'apprentissage simultané de plusieurs tâches. En pré-entraînant le modèle sur plusieurs tâches, nous nous assurons qu'il ne s'optimise pas uniquement pour une tâche prétexte, mais qu'il développe plutôt une représentation commune des images qui lui permet de traiter plusieurs tâches différentes, améliorant ainsi

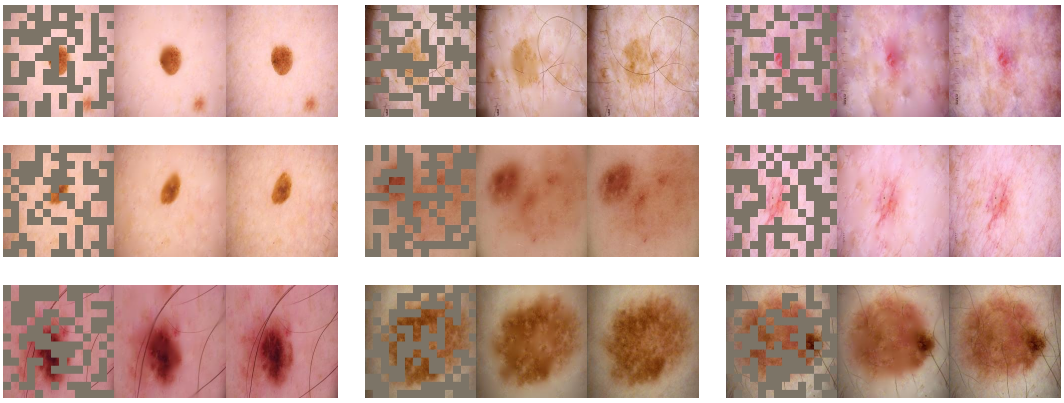

FIGURE 3.10. – **Exemples de résultats sur les images de validation d'ISIC 2020.** (à gauche) vérité terrain, (au milieu) l'image masquée, (à droite) reconstruction des patchs manquants par SJ-MAE. Le ratio de masquage est de 50 %, ne laissant que 98 patches sur 196.

les performances sur une variété de tâches ultérieures. Bien que prometteuse, notre architecture développée nécessite également plus de calculs par rapport à d'autres approches, entraînant un temps d'entraînement plus long. Des travaux futurs pourraient analyser les gains de performance lorsque le modèle est entraîné pendant des époques supplémentaires et sur d'autres jeux de données plus diversifiés, tels qu'ImageNet. Nous croyons que l'architecture SJ-MAE pourrait être utilisée pour initialiser de nombreux modèles et accélérer leurs entraînements.

3.4. Synthèse

L'apprentissage de représentations visuelles est un domaine vaste et complexe. La multitude et la diversité des modèles qui le composent montrent qu'il existe de nombreuses manières de comprendre les images sans même utiliser d'annotations. Ce chemin de recherche nous fait prendre conscience de la complexité de la vision, notamment en ce qui concerne la détection des objets, qui semble pourtant si simple à apprendre chez les animaux.

À travers nos expériences, nous avons constaté que ces approches originales permettent non seulement de pré-entraîner les modèles de manière efficace, mais aussi que leur fusion peut améliorer leur capacité à générer des embeddings performants. En offrant aux modèles différents points de vue et plusieurs objectifs, nous espérons que ces expériences inspireront la communauté scientifique à développer des modèles permettant une compréhension des images et du monde plus riche et diversifiée.

Par ailleurs, bien que nous ayons travaillé uniquement sur des images, leur application dans le domaine de la vidéo représente une opportunité d'amélioration future pour de nombreux autres domaines de recherche.

Enfin, bien que nous disposions désormais de bonnes représentations visuelles, il

nous reste à étudier un problème important : l'interprétabilité. Le chapitre suivant présentera ainsi la suite de nos travaux en ce sens, visant à donner du sens à cet espace vectoriel nouvellement généré et à mieux comprendre les modèles de classification.

4. La génération d'images, une voie vers l'interprétabilité

Sommaire

4.1. Introduction 94
4.2. Etat de l'art 95
4.2.1. Approches générales d'interprétabilité 96
4.2.2. La génération d'images de manière générale 99
4.2.2.1. AutoEncodeurs 99
4.2.2.2. AutoEncodeurs Variationnels 100
4.2.2.3. Modèles Génératifs Adversariaux 100
4.2.2.4. Modèles de Diffusion 101
4.2.3. L'interprétabilité à travers la génération d'images 102
4.2.3.1. Adversarial Latent AutoEncoders 102
4.2.3.2. StyleGAN et inversions 102
4.3. Interprétation de classifieurs par génération d'images 103
4.3.1. La génération d'images difficiles à diagnostiquer 104
4.3.2. Méthodologie 104
4.3.3. Gestion des artefacts 105
4.3.4. Images générées 105
4.3.5. Impact de l'ajout des images synthétiques sur la classification . 108
4.3.6. Analyse de l'espace latent 109
4.4. Synthèse 110

4.1. Introduction

Dans le chapitre précédent, nous avons étudié les méthodes permettant d'acquérir de bons plongements à l'aide de modèles de représentations visuelles. Toutefois, l'espace latent généré par ces approches, bien que très précis pour les classifications et les comparaisons, reste très complexe à déchiffrer et à comprendre.

Ainsi, nos derniers travaux se sont focalisés sur le problème de l'interprétabilité, un domaine extrêmement important en médecine, notamment lors de l'intégration d'approches automatiques en clinique. Selon le Centre National de Ressources Textuelles et Lexicales, la définition du verbe "interpréter" est « Expliquer, chercher à

rendre compréhensible ce qui est dense, compliqué, ambigu ». Ainsi, un modèle d'intelligence artificielle dit "interprétable" serait un modèle dont le fonctionnement, par nature dense, compliqué et ambigu, pourrait être expliqué et rendu compréhensible.

Dans le Chapitre 1, nous avons vu que, bien que les systèmes experts aient des performances moindres, leur mode de fonctionnement pouvait être compris assez facilement, car basés sur des concepts et caractéristiques bien établis par leurs concepteurs. Contrairement à ces derniers, les réseaux de neurones restent moins compris, en particulier ce qui concerne leur prise de décision, cela est du à leur fonctionnement qui repose essentiellement sur des opérations de multiplications de matrices et de fonctions d'activation dont les valeurs internes sont difficiles à comprendre.

Cette limite a ouvert une nouvelle voie pour la recherche scientifique, faisant émerger le domaine de la "XAI" (pour eXplainable AI, intelligence artificielle explicable).

Dans ce chapitre, nous étudierons ce domaine pour établir son impact, présent et futur, appliqué aux lésions cutanées. Tout d'abord, nous présenterons les premières approches générales d'interprétabilité présentes dans la littérature, avec leurs avantages et inconvénients, puis nous analyserons les innovations qui, au fil du temps, ont façonné ce domaine.

Ensuite, nous nous intéresserons tout particulièrement aux approches génératives, initialement conçues pour des objectifs créatifs, mais qui présentent de nombreux avantages pour la compréhension de l'espace latent. Bien qu'encore peu utilisées en XAI, ces techniques, regroupant la manipulation de l'espace latent, l'apprentissage de représentations décorrélées et l'inversion de modèles StyleGAN, offrent des possibilités intéressantes pour le domaine des lésions cutanées et pour l'explication des modèles d'IA.

Enfin, nous présenterons nos premiers travaux ainsi que les perspectives dans ce domaine en pleine expansion. Notre première approche constitue la première étape d'un projet plus global et se concentre principalement sur l'analyse des artefacts, la génération d'images dermoscopiques réalistes et leur impact sur la classification.

Bien que ces premières méthodes soient encore loin de fournir une explication détaillée des lésions et des modèles d'IA, elles représentent un premier pas vers de nouvelles manières de comprendre à la fois les maladies dermatologiques et les modèles qui les analysent.

4.2. Etat de l'art

Dès les premiers réseaux de neurones, plusieurs recherches se sont intéressées à la compréhension des sorties intermédiaires de ces modèles et à la possibilité que ces dernières puissent contenir de l'information utile, voire interprétable, pour représenter les images. En 2013, à titre d'exemple, Zeiler et al. [100] se sont penchés sur cette question, développant une nouvelle architecture et une méthode de visualisation des sorties intermédiaires, ces sorties étant aujourd'hui appelées des "caractéristiques". Leur approche repose essentiellement sur la visualisation des filtres appris et des couches de déconvolution. Cette méthode fut une première approche intéressante

pour la visualisation des CNN et de leur fonctionnement, car elle permettait de visualiser, dans une certaine mesure, les différentes formes détectées par le modèle.

De nombreuses approches s'en sont inspirées, tandis que d'autres ont opté pour une analyse du gradient, comme la méthode GRAD-CAM [101] et ses variantes, aujourd'hui très connues. Ces approches, tout aussi intéressantes, reposent essentiellement sur l'analyse de réseaux déjà entraînés pour la classification. Elles ont permis de mieux comprendre l'information présente dans les vecteurs intermédiaires, ainsi que leur utilité pour d'autres problèmes potentiels. Dans cette partie, nous passons en revue les différentes approches existantes dans la littérature scientifique.

4.2.1. Approches générales d'interprétabilité

Tout d'abord, il existe des méthodes de visualisation qui consistent à analyser, à posteriori, les valeurs internes des réseaux de neurones, telles que les gradients ou les filtres des couches intermédiaires appris par le réseau.

Cartes de chaleur et saliency maps Les cartes de chaleur basées sur les gradients, telles que Gradient-weighted Class Activation Mapping (Grad-CAM) [101] et ses variantes comme Grad-CAM++ [102] et Score-CAM [103], constituent une méthode puissante pour interpréter les modèles. Grad-CAM utilise les gradients d'une classe cible par rapport aux cartes de caractéristiques d'une couche convolutive pour produire une carte de chaleur visuelle qui indique les régions importantes pour la prédiction d'une classe spécifique. Cette technique est particulièrement utile pour la visualisation des décisions de modèles dans des applications telles que la reconnaissance d'images, où elle peut identifier les zones de l'image qui ont contribué le plus à la classification. Cependant, Grad-CAM peut être limité par sa dépendance à des couches spécifiques et son incapacité à toujours capturer les contributions subtiles des caractéristiques à plus petite échelle. De plus, il peut être difficile d'interpréter correctement les cartes de chaleur sans une compréhension approfondie du modèle sous-jacent. Un exemple de ce genre de carte de chaleur est présenté en Figure 4.1.

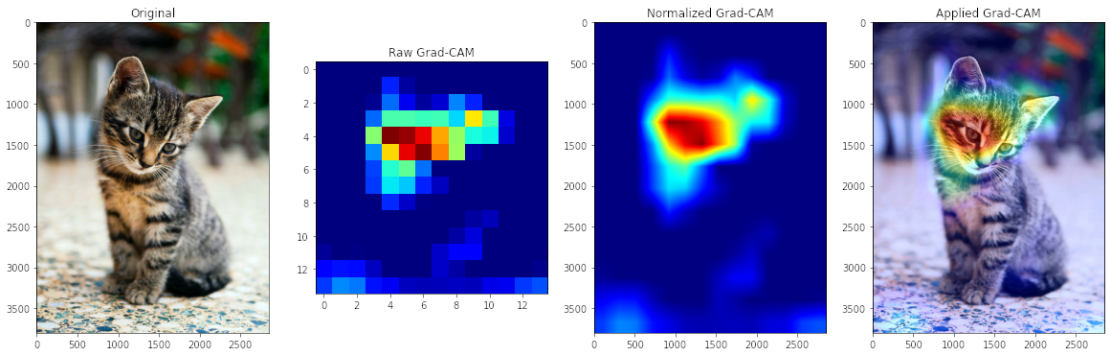

FIGURE 4.1. – **Exemple de carte de chaleur obtenue avec Grad-CAM.** Ici, les zones rouges représentent les endroits où le modèle a un gradient élevé pour la catégorie "chat".

Les inconvénients de ces approches se manifestent également de manière plus marquée pour images dermoscopiques, rendant cette méthode très limitée pour notre cas d'usage. En effet, les gradients montrent généralement une forte activation sur les lésions elles-mêmes, ce qui n'apporte pas réellement d'informations supplémentaires utiles au diagnostic. Plusieurs exemples de cartes de chaleur sur des images dermoscopiques sont présentés à la Figure 4.2. Plusieurs travaux ont incorporé ces cartes de chaleurs dans des architectures plus grandes [104], mais l'apport en interprétabilité reste limité.

Il est à noter que ces méthodes restent très utiles dans de nombreux domaines où l'emplacement des gradients dans l'image est important. Par exemple, de nombreux travaux permettent d'analyser l'emplacement de tumeur à l'aide de ces méthodes [105] [106].

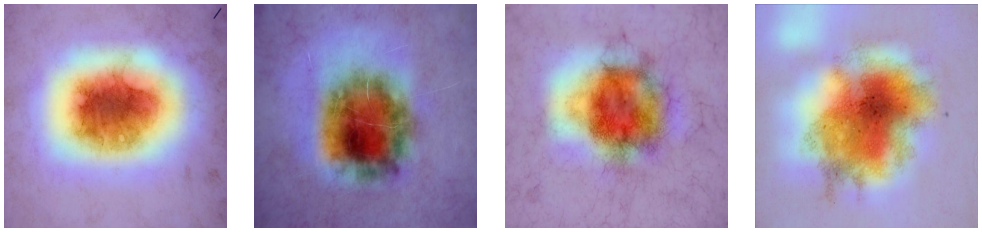

FIGURE 4.2. – **Quatre exemples de cartes de chaleur sur des images dermoscopiques.** Le rouge représente les zones à forte activation de gradient pour la catégorie 'mélanome'. On remarque que les cartes de chaleur se focalisent effectivement sur les lésions, ce qui montre que le modèle se concentre au bon endroit pour réaliser le diagnostic. Toutefois, cette information n'est pas directement utile aux dermatologues lors de l'analyse des lésions.

Visualisation des filtres et des couches intermédiaires Cette technique, présentée notamment dans un article de 2014 [107], permet de visualiser les activations des couches intermédiaires d'un réseau de neurones, offrant un aperçu des caractéristiques que le modèle extrait à différents niveaux de son architecture. Par exemple, les premières couches peuvent capturer des éléments simples comme les bords, tandis que les couches plus profondes identifient des caractéristiques plus complexes. Cette méthode aide à comprendre comment le modèle construit progressivement sa compréhension des images. La Figure 4.3 présente quelques exemples de filtres appris par un CNN sur plusieurs classes différentes. Bien qu'intéressante, cette approche reste également limitée, car les filtres appris peuvent être visuellement très difficiles à analyser.

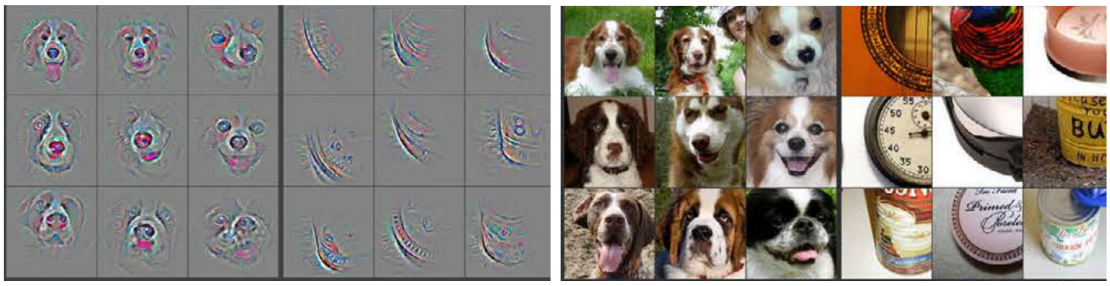

FIGURE 4.3. – **Visualisation de filtres appris par un CNN.** On constate que le modèle a appris à reconnaître des formes complexes dans le cas d'images de chiens, par exemple. *Image provenant de [107]*

Dans le cadre des lésions cutanées, la compréhension des filtres est également très limitée et difficile à interpréter, d'autant plus que les lésions peuvent présenter des signes très subtils de malignité qui ne sont pas forcément représentés au sein des filtres.

Local Interpretable Model-agnostic Explanations (LIME) LIME [108] est une technique qui explique les prédictions des modèles sur une base de données par instance. En perturbant l'entrée et en observant les changements dans les sorties, LIME crée un modèle simple, localement valide, pour approcher le comportement du modèle original autour d'une prédiction spécifique. Dans le domaine des images, cette approche revient également à générer une sorte de carte de chaleur. Dans le domaine des lésions cutanées, quelques travaux ont utilisé cette approche, permettant la visualisation de cartes de chaleur [109, 110], différentes des cartes issues de Grad-CAM, offrant une couche supplémentaire d'interprétabilité.

SHapley Additive exPlanations (SHAP) SHAP [111] est basé sur la théorie des jeux et attribue à chaque caractéristique une importance en termes d'impact sur la différence entre la prédiction actuelle et la prédiction moyenne. SHAP fournit une mesure cohérente et précise de l'importance des caractéristiques, ce qui facilite grandement l'interprétation des modèles, notamment lorsque ces derniers utilisent une liste de caractéristiques pour réaliser les prédictions. Dans le cas du mélanome, par exemple, cette approche peut s'avérer très utile lorsque les modèles utilisent d'autres informations en plus des images. Plusieurs articles basés sur des informations multimodales combinant images et informations patientes ont pu utiliser les valeurs SHAP afin d'analyser les caractéristiques les plus impactantes pour la classification [112]. Certaines approches fusionnant les images avec d'autres caractéristiques expertes, comme l'asymétrie des lésions, ont pu utiliser les valeurs SHAP dans le but de déterminer, pour chaque lésion, dans quelle mesure ces caractéristiques ont été déterminantes lors de la classification [43]. Bien que les valeurs SHAP soient très efficaces, cela implique un coût computationnel plus élevé et un temps d'analyse plus long.

Visualisation des Décisions de Modèle Enfin, la génération de contre-exemples et d'exemples synthétiques aide à identifier les limites et les comportements de généralisation des CNNs. En manipulant des caractéristiques spécifiques dans les images générées, les chercheurs peuvent observer directement l'impact de ces caractéristiques sur les prédictions du modèle [113].

Ces différentes approches sont très intéressantes, mais leur usage reste limité dans le domaine des images dermoscopiques. L'approche que nous avons voulu étudier a été la génération d'images à des fin d'interprétabilité, car cette dernière semble la plus prometteuse à inclure dans un contexte médical. Nous avons donc étudié à la fois les méthodes d'interprétabilité et de génération d'images afin de continuer nos travaux. Dans la partie suivante, nous nous intéressons donc au domaine de la génération d'images.

4.2.2. La génération d'images de manière générale

En parallèle à cela, la génération d'images a connu de nombreuses avancées ces dernières années, allant des AutoEncodeurs (AEs) [56] aux Réseaux Antagonistes Génératifs (GANs) [114], et plus récemment, aux modèles basés sur la diffusion [115] [116]. Ces avancées, que nous allons présenter ici, ont conduit à la création d'images de plus en plus réalistes, avec un contrôle de plus en plus précis.

4.2.2.1. AutoEncodeurs

Les premiers modèles de génération d'images sont les AutoEncodeurs (AEs) [56], présentés pour la première fois en 1992. Ces réseaux de neurones sont composés d'un encodeur et d'un décodeur, permettant la compression et la reconstruction des données. L'encodeur $f_{\text{enc}}(\mathbf{x};\theta_{\text{enc}})$ projette une image d'entrée $\mathbf{x} \in \mathbb{R}^n$ dans un espace latent de dimension inférieure $\mathbf{z} \in \mathbb{R}^m$:

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x};\theta_{\text{enc}})$$

Les images passent par un goulot d'étranglement où l'information est compressée, puis à partir de cette information compressée, le décodeur $f_{\text{dec}}(\mathbf{z};\theta_{\text{dec}})$ reconstruit l'image originale $\hat{\mathbf{x}} \in \mathbb{R}^n$:

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z};\theta_{\text{dec}})$$

Cependant, ces modèles restent rudimentaires et présentent plusieurs problèmes, comme la difficulté à capturer la complexité des données, la perte de détails importants lors de la reconstruction, et une tendance à produire des résultats souvent flous ou peu réalistes. De plus, ces modèles ont du mal à se généraliser à des données non vues pendant l'entraînement, rendant la génération de nouvelles images très difficile. Ces problèmes sont issus de leur fonction de perte qui repose sur la minimisation de la différence entre l'image d'entrée $\mathbf{x}$ et l'image reconstruite $\hat{\mathbf{x}}$, souvent mesurée par l'erreur quadratique moyenne (MSE) :

$$\min_{\theta_{\text{enc}},\theta_{\text{dec}}} \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[\|\mathbf{x} - f_{\text{dec}}(f_{\text{enc}}(\mathbf{x};\theta_{\text{enc}});\theta_{\text{dec}})\|_2^2\right]$$

4.2.2.2. AutoEncodeurs Variationnels

En 2013, les Variational Autoencoders (VAEs) [71] introduisent une dimension probabiliste qui améliore la génération de nouvelles images en rendant l'espace latent plus structuré et continu. Dans un VAE, l'encodeur ne produit pas directement un code latent $\mathbf{z}$, mais plutôt les paramètres d'une distribution probabiliste $q_\phi(\mathbf{z}|\mathbf{x})$, généralement une distribution gaussienne avec une moyenne μ et une variance σ^2 :

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z};\mu(\mathbf{x}),\sigma^2(\mathbf{x}))$$

L'espace latent devient alors une variable aléatoire, et le décodeur génère une reconstruction $\hat{\mathbf{x}}$ à partir d'un échantillon $\mathbf{z}$ tiré de cette distribution :

$$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$$

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z};\theta_{\text{dec}})$$

Cette approche permet une meilleure modélisation de la distribution des données en maximisant une fonction objectif qui inclut une divergence de Kullback-Leibler (KL) entre $q_\phi(\mathbf{z}|\mathbf{x})$ et une distribution à priori $p(\mathbf{z})$ généralement choisie comme une distribution normale standard $\mathcal{N}(0,I)$:

$$\mathcal{L}(\mathbf{x},\hat{\mathbf{x}}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \text{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\,\|\,p(\mathbf{z})\right)$$

Cette approche facilite également l'exploration et la manipulation de l'espace latent pour diverses applications, allant de la synthèse d'images à l'amélioration de la compréhension des caractéristiques latentes des données.

4.2.2.3. Modèles Génératifs Adversariaux

Cependant, le domaine de la génération d'images a radicalement changé en 2014, avec le développement des modèles génératifs adversariaux (GANs) [117]. Cette nouvelle approche utilise deux modèles distincts en compétition : un générateur G produisant des images, et un discriminateur D les classant entre "réelles" et "générées". Le générateur prend un vecteur aléatoire $\mathbf{z}$ échantillonné à partir d'une distribution $p_{\mathbf{z}}(\mathbf{z})$ (par exemple, une distribution gaussienne) et génère une image $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = G(\mathbf{z};\theta_G)$$

Le discriminateur $D(\mathbf{x};\theta_D)$ estime la probabilité qu'une image donnée $\mathbf{x}$ soit réelle (provenant des données d'entraînement) plutôt que générée par le générateur :

$$D(\mathbf{x};\theta_D) \rightarrow [0,1]$$

L'entraînement des GANs se fait en résolvant un problème d'optimisation min-max où le générateur tente de minimiser la capacité du discriminateur à distinguer les images réelles des images générées, tandis que le discriminateur tente de maximiser cette capacité. La fonction de coût pour ce jeu est donnée par :

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

Cette approche innovante a immédiatement connu de très bons résultats, générant des images réalistes. Les GANs ont dominé le domaine de la génération d'images entre 2014 et 2020, inspirant de nombreuses variantes, comme les Conditional GANs [118] et les CycleGANs [119], permettant de conditionner la génération sur des données spécifiques et de modifier des images entre différents domaines.

Cependant, malgré leurs succès, les GANs présentent des limitations importantes, notamment leur incapacité à recréer exactement une image donnée et la difficulté à manipuler leur espace latent. En effet, en apprenant uniquement à tromper un modèle discriminant à base d'une distribution aléatoire, les générateurs sont optimisés pour générer des images réalistes sans réel contrôle sur leur espace latent. Ce problème était déjà connu à l'époque, et nous verrons dans la section suivante les variantes développées pour y remédier.

4.2.2.4. Modèles de Diffusion

En 2020, le monopole des GANs a légèrement décliné avec le regain de popularité des modèles de diffusion. Initialement conçus à partir de 2015 par Sohl-Dickstein et al. [120], ces modèles reposent sur un processus de bruitage progressif des images, suivi de l'apprentissage du débruitage inverse pour générer de nouvelles instances. Bien que peu populaires à leurs débuts, ces approches sont devenues très en vogue en 2020 avec la publication des modèles Denoising Diffusion Probabilistic Models (DDPM) [121] puis de Stable Diffusion [115], qui ont montré des résultats impressionnants en termes de qualité et de diversité des images générées.

Contrairement aux GANs, les modèles de diffusion visent généralement à apprendre la distribution sous-jacente des données, souvent en utilisant une fonction de perte qui inclut des termes pour la reconstruction des données, comme la fonction de perte des modèles de diffusion conditionnelle de débruitage :

$$L_{\text{LDM}} := \mathbb{E}_{(x,y),\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\right\|_2^2\right]$$

Ici, L_{LDM} calcule l'espérance de la norme L2 au carré de la différence entre le bruit réel ϵ et le bruit prédit par le modèle ϵ_θ, conditionnée sur la variable latente z_t, l'instant t, et les étiquettes transformées $\tau_\theta(y)$. Ces modifications se sont avérées efficaces pour générer des images de meilleure qualité bien qu'étant très coûteuses en terme d'énergie.

4.2.3. L'interprétabilité à travers la génération d'images

Ces modèles ont ainsi permis de générer des images très réalistes et variées, mais bien plus que la simple génération d'images, nous nous sommes demandé s'il n'était pas possible d'utiliser ces modèles génératifs pour d'autres tâches, notamment celle de l'interprétabilité. En effet, certains modèles projettent les images dans un espace latent, plus petit, qu'il serait probablement possible d'analyser afin de visualiser les différents types d'images que le modèle est capable de générer. En suivant cette logique, plusieurs approches se sont intéressées à la génération d'images tout en conservant la possibilité de garder ces vecteurs issus des images, et également de les comprendre et de les manipuler.

4.2.3.1. Adversarial Latent AutoEncoders

Tout d'abord, les Adversarial Latent Autoencoders (ALAE) [59] (2020) combinent les principes des autoencodeurs et des GANs pour produire des représentations latentes de haute qualité à partir des données d'entrée. L'autoencodeur encode les données dans un espace latent compact, tandis qu'un réseau antagoniste veille à ce que cette distribution suive une distribution prioritaire désirée, souvent gaussienne. Cette configuration permet non seulement de générer des données réalistes à partir de l'espace latent, mais aussi de comprendre les facteurs de variation des données originales.

Les techniques de visualisation des représentations latentes apprises par les ALAE peuvent être mises en œuvre de plusieurs manières. L'interpolation dans l'espace latent consiste à interpoler entre deux points de cet espace et à générer les images correspondantes, ce qui permet d'observer l'évolution continue des caractéristiques visuelles et de révéler comment le modèle perçoit les transformations entre différents états ou objets. La manipulation dirigée implique l'ajustement délibéré de certaines dimensions de l'espace latent. En observant les modifications dans les images générées, les chercheurs peuvent identifier quelles dimensions contrôlent des caractéristiques spécifiques dans les données visuelles. Dans l'article original entraîné sur la base de données FFHQ, composée de visages [122], le modèle est capable d'ajouter des lunettes à des personnes, de les faire sourire, de leur ajouter une moustache, etc., uniquement à partir de translations dans l'espace latent, ce qui est déjà très impressionnant. Le désavantage de cette approche est que les chercheurs ont dû trouver ces vecteurs directeurs de manière relativement manuelle.

4.2.3.2. StyleGAN et inversions

En parallèle, StyleGAN [122] et ses itérations ultérieures, StyleGAN2 [123] et StyleGAN3 [124], ont introduit des techniques puissantes pour l'édition et la manipulation d'images, permettant un contrôle précis sur les styles et les éléments présents dans les images. Ces modèles sont particulièrement pertinents pour notre travail car ils offrent des outils robustes pour explorer et manipuler l'espace latent des images. De manière similaire aux ALAE, StyleGAN permet de conditionner et de modifier les images en

fonction de vecteurs choisis, rendant ainsi possible l'édition des images par la modification des valeurs de l'espace latent. En utilisant ces techniques, nous pouvons augmenter l'interprétabilité des sorties des modèles d'apprentissage, en permettant une visualisation et une compréhension plus approfondie des transformations et des caractéristiques visuelles capturées par les réseaux neuronaux.

Ainsi, la capacité des ALAE et des modèles comme StyleGAN à fournir une visualisation claire et manipulable de l'espace latent ouvre des perspectives dans le domaine des lésions cutanées, mais également dans d'autres domaines tels que la bioinformatique pour la découverte de biomarqueurs, la reconnaissance faciale avancée et le développement de contenus créatifs personnalisés. Comprendre les représentations latentes permet de raffiner et d'ajuster le processus d'apprentissage pour obtenir des résultats plus précis et pertinents.

Dans le cas des lésions cutanées, nous pouvons donc conceptualiser des approches permettant la génération d'images conditionnées par du texte ou par un emplacement dans l'espace latent, par exemple. De cette manière, nous espérons détecter différents clusters représentant des caractéristiques visuelles spécifiques aux lésions, afin de découvrir, idéalement, les liens entre certains éléments des lésions et la présence d'un mélanome.

Dans les travaux que nous présentons dans la partie suivante, nous utiliserons les modèles de diffusion, encore peu étudiés dans le domaine des lésions cutanées, qui présentent une capacité de génération considérable. De plus, les annotations généralement utilisées pour entraîner les modèles génératifs se basent uniquement sur la classe de l'image, ce qui peut poser problème en cas d'incertitude diagnostique. Nous employons donc, dans nos approches, des méthodes différentes pour conditionner le modèle génératif, que nous expliquerons dans la section suivante.

4.3. Interprétation de classifieurs par génération d'images

Les innovations en matière de génération d'images nous ont donc inspirés à développer ces approches dans le cadre des lésions cutanées. Bien que plusieurs travaux abordent la génération d'images dermoscopiques, peu se penchent sur l'impact potentiel que ces modèles pourraient avoir en termes d'interprétabilité au niveau clinique, et encore moins utilisent les modèles de diffusion dans le domaine des images dermoscopiques.

De manière générale, les articles utilisant des approches génératives les emploient comme forme d'augmentation de données lors de l'entraînement afin d'accroître virtuellement le nombre d'images vues par le modèle [125, 126]. Ces études semblent montrer que l'utilisation d'images générées par intelligence artificielle permet d'améliorer les performances des modèles de classification subséquents. Cependant, les articles présentant de tels résultats sont encore rares, et la littérature ne semble pas fournir de conclusions claires sur la capacité et les limites de ces modèles génératifs. Néanmoins, ces modèles restent utiles, et il est envisageable de générer des images

artificielles pour visualiser l'espace latent appris par les réseaux de neurones.

Dans cette section, nous présentons nos premières approches dans le domaine. Nous avons d'abord conçu un modèle générant des images sans artefacts, de sorte que les images synthétiques soient aisément analysables. Ensuite, nous avons cherché à générer des images pour lesquelles les classifieurs réalisaient de mauvaises prédictions, afin de mieux comprendre les caractéristiques que ces derniers n'arrivaient pas à détecter dans certains mélanomes. Nous générons ces images, puis les ajoutons à la base de données afin de réentraîner un CNN et vérifier si cette approche permet une amélioration des performances pour la tâche de classification.

À terme, l'idée générale de notre approche repose sur deux étapes. Premièrement, encoder les images dans un espace latent continu où les composantes principales correspondraient aux caractéristiques visuelles des lésions. Ensuite, analyser cet espace latent en générant des images à différents endroits de celui-ci, afin de situer certaines caractéristiques visuelles dans cet espace. Cette approche serait particulièrement utile pour les lésions cutanées, en rendant possible la visualisation des caractéristiques des mélanomes et des nævi qui ne sont pas visibles avec les méthodes précédentes.

4.3.1. La génération d'images difficiles à diagnostiquer

Ainsi, nous avons d'abord voulu vérifier la capacité des images synthétiques à améliorer les performances sur la tâche de classification lorsque ces dernières sont données en entrées à un classifieur, en plus des images originales. Certains travaux précédents ont noté une amélioration des performances à l'aide des images générées, mais les articles présentant ces résultats restent rare, que ce soit dans le domaine des lésions cutanées que dans d'autres domaines.

4.3.2. Méthodologie

Nous présentons ici notre méthodologie, dont la majeure partie reste la même dans les approches suivantes. Nous conserverons, au fil des expériences, la même architecture globale, à laquelle nous ajoutons progressivement différentes améliorations en fonction de nos travaux.

Pour ce faire, nous avons utilisé le modèle Stable Diffusion [115] pré-entraîné sur la base de données LAION [127], que nous avons fine-tuné à l'aide de notre propre base de données dermatologique. Afin d'accélérer l'entraînement, nous avons choisi d'affiner le modèle sur l'espace latent plutôt que d'apprendre l'ensemble du processus de bruitage. Le fine-tuning du modèle Latent Diffusion a été effectué sur 50 époques en utilisant les paramètres par défaut décrits dans l'article original. Pour conditionner la génération, le texte d'entrée (appelé prompt) indique si l'image représente un mélanome ou non, ainsi que la prédiction obtenue par un CNN de classification. Si l'image est un mélanome, nous ajoutons le mot "melanoma" au prompt ; sinon, nous utilisons "naevus", suivi de "correct" si la classification est correcte ou "incorrect" si elle ne l'est pas. Pour garantir la génération d'images sans artefacts, nous avons également intégré des annotations appropriées à l'aide d'annotations manuellement étiquettées

dans d'autres travaux présents dans la littérature [128]. Cette architecture générative permet une approche nouvelle pour comprendre la performance des classificateurs. Par exemple, générer des images représentant des cas ambigus, comme des lésions prédites à 50 % naevus et 50 % mélanome, offre la possibilité d'analyser les défis auxquels sont confrontés les classificateurs et les dermatologues lors du diagnostic du mélanome.

4.3.3. Gestion des artefacts

Nos premiers travaux se sont donc basés sur la génération d'images dermoscopiques à l'aide de modèles de diffusion. Plus précisément, nous avons commencé par générer des images sans artefacts. Dans les images dermoscopiques, les artefacts sont des éléments qui empêchent la bonne visualisation des lésions ou qui ne sont pas utiles à leur analyse. Les artefacts les plus courants sont les poils devant la lésion, les bulles d'air présentes sur la lésion lorsque le dermatoscope nécessite un gel spécial pour la prise de photo, ainsi que des taches d'encre, des règles et des patchs collants parfois utilisés par les dermatologues pour la calibration, mais qui peuvent nuire au processus de génération d'images.

Ces artefacts constituent par ailleurs un problème majeur dans la base de données ISIC 2020, qui contient une part importante d'images avec des artefacts.

Pour éviter de générer ces éléments, nous avons entraîné un modèle sur une base de données existante, annotée par des dermatologues, indiquant pour chaque lésion la présence ou l'absence d'artefacts parmi plusieurs catégories telles que "poils", "encre" et "bulles d'air". Cette base de données, présentée en 2022 par [128], permet ainsi d'avoir une annotation précise pour chaque image concernant les artefacts.

4.3.4. Images générées

Une fois entraîné, notre modèle a pu générer des images comportant moins d'artefacts que les images originales. Nous présentons plusieurs exemples d'images générées à l'aide du modèle de diffusion dans les Figures 4.4, 4.5, 4.6 et 4.7. La Figure 4.4 présente des mélanomes générés pour lesquels le modèle avait la condition de générer un mélanome bien prédit par le CNN. La Figure 4.5 montre des exemples de mélanomes qui seraient mal prédit, la Figure 4.6 des images de naevus bien prédits et enfin la Figure 4.7 montre des naevus générés dont le CNN classifierait comme "mélanome".

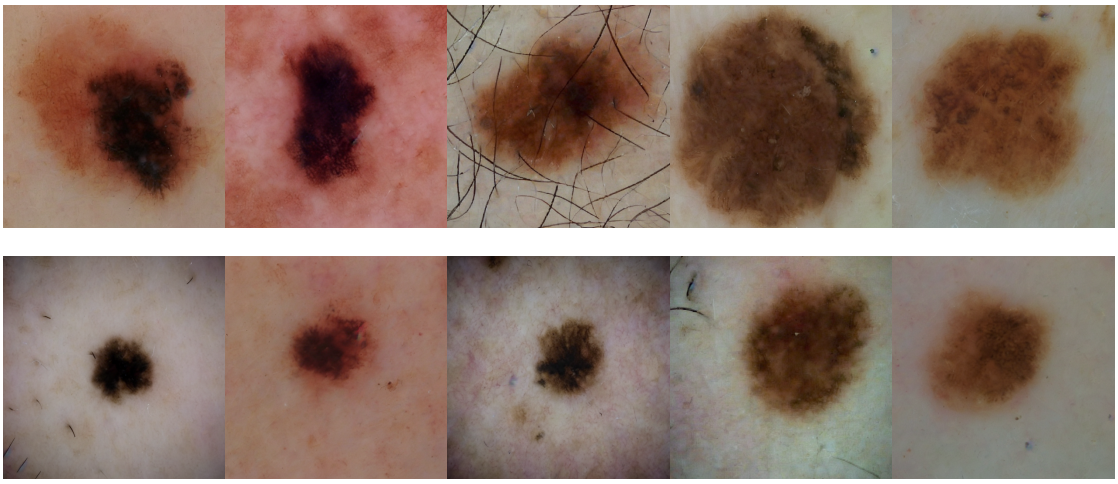

FIGURE 4.4. – **Images générées par le modèle avec la condition de générer un mélanome bien prédit par le classifieur.**

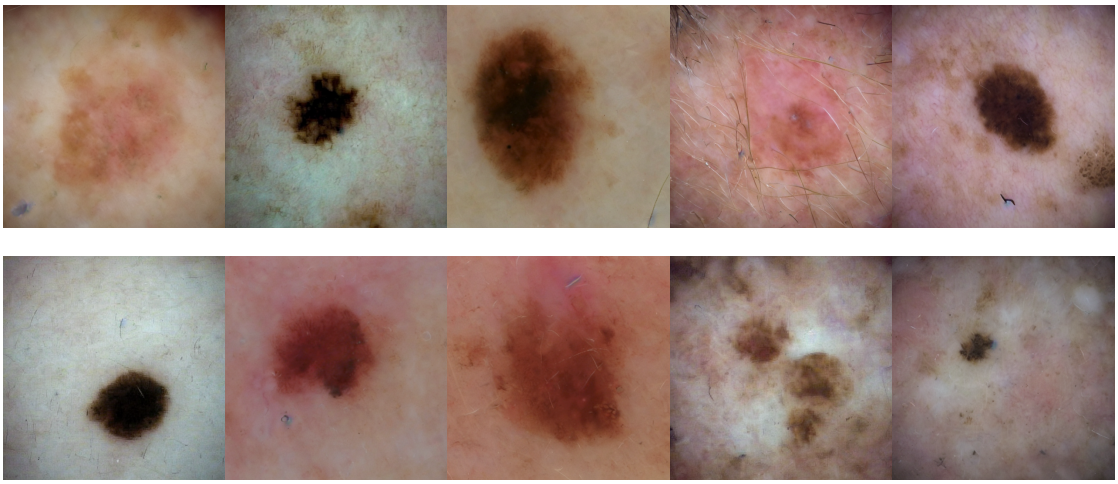

FIGURE 4.5. – **Images générées par le modèle avec la condition de générer un mélanome mal prédit par le classifieur.**

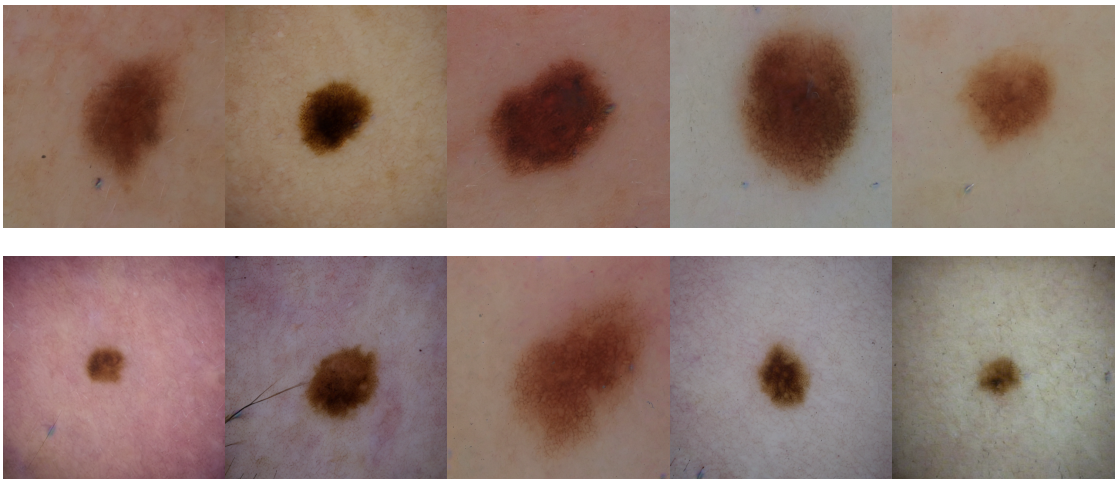

FIGURE 4.6. – **Images générées par le modèle avec la condition de générer un naevus bien prédit par le classifieur.**

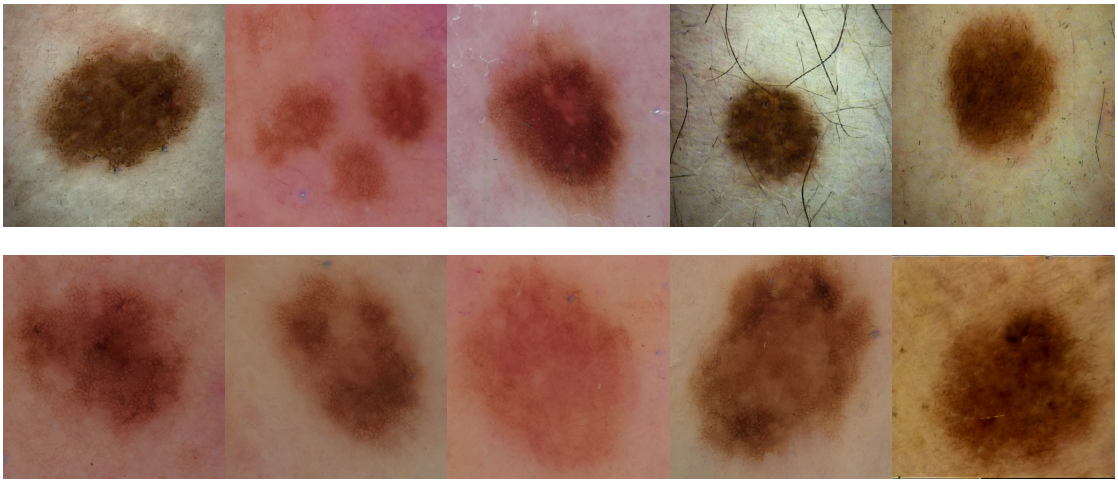

FIGURE 4.7. – **Images générées par le modèle avec la condition de générer un nevus mal prédit par le classifieur.**

Nous remarquons alors que, même en présentant une belle diversité d'images, celles-ci sont également fortement impactées par la description donnée en entrée au modèle. Les naevus bien prédits présentent pour la plupart des formes plus circulaires, ordonnées et également plus petites, tandis que les autres images présentent des aspects plus atypiques. Cela fait sens, car les autres catégories comportent les mélanomes bien prédits, ceux confondus avec des naevus ainsi que les naevus confondus avec des mélanomes. Ces trois catégories sont donc plus ou moins directement liées à la catégorie mélanome, ce qui se répercute sur les images générées. Les mélanomes bien prédits présentent une taille plus importante, des bordures et des couleurs désordonnées, qui sont des marqueurs importants des mélanomes.

Les catégories représentant les erreurs du CNN sont elles très intéressantes et plus complexes. Les mélanomes mal prédits, par exemple, présentent une diversité accrue, notamment au niveau de la peau, cette dernière affichant plus fréquemment des aspects, couleurs et formes désordonnées. Nous remarquons également que certaines lésions sont plutôt claires, d'autres foncées mais avec une forme arrondie. Ainsi, ces images présentent à la fois des caractéristiques des mélanomes, mais également certaines des naevus. Ces différences au niveau de la génération d'images peuvent probablement être la conséquence des limites du modèle de classification. Ce dernier, bien que capable de discerner les caractéristiques des mélanomes et des naevus, se retrouve en difficulté lorsque les lésions présentent plusieurs critères discordants.

Les naevus mal classés, eux, présentent également ces aspects, avec des formes et couleurs évoquant les deux catégories. Enfin, on aperçoit également que les images générées représentant les erreurs comportent des images avec deux lésions, comme on peut le voir sur la Figure 4.5 et la Figure 4.7, ce qui est beaucoup plus rare dans les images générées représentant les images bien classées.

4.3.5. Impact de l'ajout des images synthétiques sur la classification

Nous avons ensuite ré-entraîné un CNN avec les nouvelles images synthétiques afin de vérifier si cet ajout améliorait les performances de classification. Les résultats sont présentés dans la Figure 4.1.

Images réelles	Images synthétiques	Poids de classes	AUC	BACC
✓	×	×	0.859	0.768
✓	×	✓	0.868	0.796
✓	✓	✓	**0.878**	**0.816**

TABLEAU 4.1. – **Comparaison des performances lors de l'ajout d'un poids de classe et lors de l'utilisation de données synthétiques.** Les meilleures valeurs par métrique sont présentées en gras pour chaque ligne. L'utilisation de données synthétiques montre une performance supérieure.

Nous avons comparé les résultats obtenus avec un CNN sans pondération des classes, un autre avec pondération des classes, et enfin un CNN utilisant des images synthétiques en plus. Les résultats montrent une légère amélioration avec la pondération des classes, ce qui est attendu. De plus, l'ajout d'images synthétiques a également conduit à une amélioration des performances. Ces résultats indiquent que les images synthétiques peuvent contribuer à de meilleurs résultats pour l'entraînement des CNN. De manière similaire aux approches de 'bagging', où plusieurs classifieurs sont entraînés pour la même tâche et leurs prédictions combinées pour produire une prédiction finale, l'intégration de modèles génératifs dans l'architecture d'entraînement introduit une diversité qui peut offrir une compréhension différente des images, améliorant ainsi les performances finales.

4.3.6. Analyse de l'espace latent

En plus de la génération d'images, nous avons souhaité visualiser la localisation des images dans l'espace latent appris par le modèle de classification. Cette méthode, couramment utilisée en pratique, permet de visualiser les différences et similarités entre les lésions. En clinique, elle peut également montrer au dermatologue les lésions similaires à une lésion donnée, offrant ainsi un support supplémentaire au diagnostic. À cet effet, nous avons développé une application permettant la visualisation des images dans l'espace latent, présentée à la Figure 4.8. Plusieurs options sont disponibles, comme la réduction de l'espace à deux dimensions à l'aide d'algorithmes de réduction de dimensions tels que UMAP [129], ou encore la possibilité de choisir différentes combinaisons de dimensions projetées en deux dimensions.

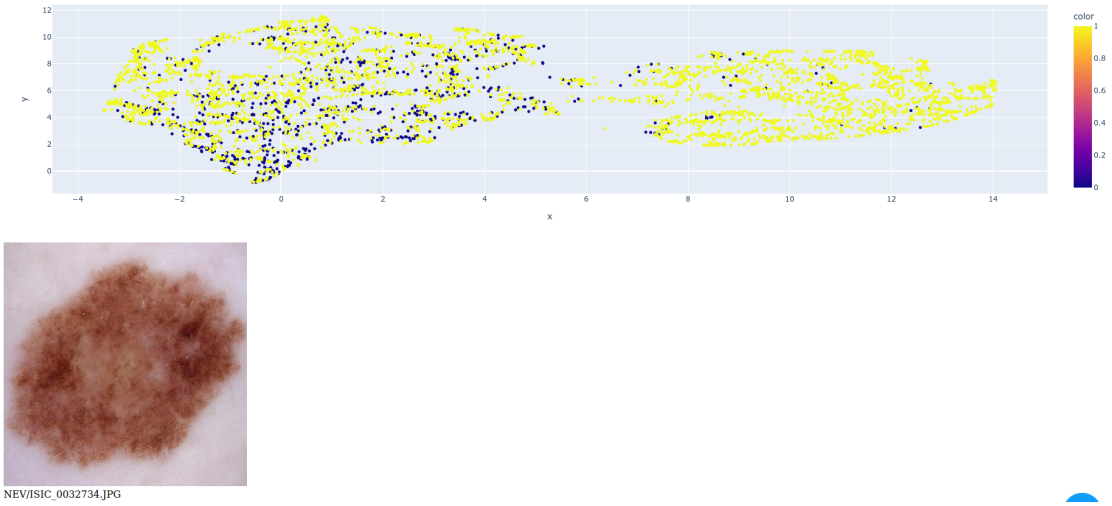

FIGURE 4.8. – **Projection des images d'ISIC 2020 en deux dimensions à l'aide des vecteurs appris par classification. Les points jaunes sont des naevus et les points violets des mélanomes. Il est possible de visualiser l'image choisie en bas de l'écran.** Nous remarquons différents clusters, présentant chacun des caractéristiques visuelles propres.

Cette visualisation permet de comprendre comment le modèle interprète les images. Si certaines caractéristiques visibles dans les clusters correspondent bien à des critères utiles pour la classification, d'autres sont plus arbitraires, comme des clusters basés sur la colorimétrie de l'image et de l'environnement. En utilisant différents modèles, nous pouvons observer l'apparition de clusters variés. Dans notre cas, après une analyse de l'espace latent, les caractéristiques les plus distinctes étaient la colorimétrie de l'image, qui forme les deux principaux clusters de la figure 4.8. D'autres clusters présentaient également des caractéristiques différentes, comme la taille de la lésion et la présence de taches au sein de la lésion. Ainsi, cette méthode peut également être utile en clinique pour aider les dermatologues, en leur présentant des images similaires avec le diagnostic correspondant, ou en visualisant la localisation de la lésion dans l'espace latent, et en comparant celle-ci avec l'emplacement des mélanomes.

4.4. Synthèse

Que ce soit pour l'interprétabilité ou la génération d'images, ces deux domaines demeurent des problèmes ouverts dans le monde de la recherche, avec chaque approche comportant son lot d'avantages et d'inconvénients qu'il est nécessaire de considérer avant toute application pratique. Dans cette section, nous avons illustré à quel point ces deux domaines sont à la fois denses et vastes.

La génération d'images voit une multitude de nouvelles approches chaque année, initialement inspirées par les AutoEncodeurs, puis les GANs, et enfin les modèles de diffusion. Aujourd'hui, ces derniers sont les plus étudiés, ce qui justifie notre choix pour nos premiers travaux.

Parallèlement, l'interprétabilité reste un domaine extrêmement complexe et difficile à définir. À travers ce chapitre, nous avons étudié les différentes approches actuelles qui tentent de comprendre les réseaux de neurones, d'abord à l'aide de cartes de chaleur, de visualisation de filtres, puis de modèles spécialisés dans l'analyse de prédictions, et enfin à la génération d'images interprétables. Malgré de nombreux progrès, il n'existe toujours pas de solution suffisante et uniforme pour toutes les tâches d'interprétabilité. Le domaine en lui-même est une tâche complexe, et la compréhension totale des modèles de manière simple pour les humains ne semble pas avoir de solution évidente.

Appliqués à la dermatologie, nous avons pu voir à travers nos premiers travaux qu'il était possible d'analyser les lésions et leurs caractéristiques à l'aide de modèles génératifs. En générant des images pour lesquelles nous savons que le modèle se trompera de classification, nous pouvons mieux connaître les failles des modèles de classification et analyser les caractéristiques difficiles à détecter pour les classifieurs.

Notre premier travail sur les modèles génératifs pour l'interprétation des classifieurs révèle plusieurs avantages, notamment une manière originale de comprendre les prédictions des modèles et de visualiser les caractéristiques des images difficiles à classer pour les modèles de classification. Dans les cas où les données sont rares, ces modèles génératifs peuvent permettre de créer de nouvelles images présentant les caractéristiques atypiques de la classe minoritaire et les erreurs des classifieurs, améliorant ainsi la compréhension des prédictions des modèles.

Notre expérience actuelle s'est principalement concentrée sur l'utilisation de modèles génératifs pour l'augmentation de données dans un cadre théorique qui met en évidence non seulement les futures recherches potentielles intégrant de manière plus intime les modèles génératifs et de classification, telles que dans des configurations d'ensemble, mais aussi démontre des applications pratiques. Plus précisément, notre modèle a généré efficacement des images réalistes conditionnées aux prédictions d'un classifieur, améliorant les performances de ce dernier avec les données supplémentaires générées. Cette approche montre des promesses pour résoudre les problèmes médicaux où la curation des données est difficile. Cependant, il existe encore des limitations dans ces approches, telles que des coûts computationnels élevés, qui peuvent affecter leur plus large applicabilité et leurs performances. De plus amples recherches sur ces questions ont le potentiel d'améliorer les performances des classifieurs lors de

l'utilisation de générateurs dans des tâches de classification.

Enfin, les travaux que nous avons présentés ici sont le fruit des derniers efforts réalisés durant cette thèse, et constituent le début d'un long chemin vers la compréhension des modèles de réseaux de neurones. Toutefois, nous espérons que ces premiers travaux permettront et encourageront de nouvelles approches futures dans ce domaine, afin que les récents modèles en génération d'images et en interprétabilité puissent encore apporter leurs bienfaits à la médecine.

Conclusions

Le mélanome, cancer cutané parmi les plus graves, constitue un défi majeur de santé publique mondiale en raison de sa mortalité croissante. La détection précoce de ce cancer est essentielle pour limiter son impact, mais elle reste confrontée à de nombreux défis. Bien que les technologies d'acquisition, telles que la génération de modèles 3D, aient connu des avancées significatives, l'amélioration des outils de diagnostic demeure indispensable. En France, le manque de dermatologues complique l'analyse systématique de la population, augmentant le risque de cancers non détectés. Pour relever ces défis, cette thèse s'est concentrée sur le développement de solutions innovantes destinées à assister les dermatologues, afin de rendre le diagnostic plus rapide, précis et accessible.

Les récents progrès de l'intelligence artificielle, notamment dans la classification, offrent un potentiel immense pour transformer le diagnostic médical. Cependant, ces approches, bien qu'efficaces, présentent encore des limites, notamment en termes d'interprétabilité et d'adaptation à des problématiques spécifiques comme le mélanome. Dans ce contexte, nos travaux se sont focalisés sur l'étude et l'exploitation du concept du Vilain Petit Canard. Ce principe, basé sur la comparaison intra-patient des lésions cutanées, était jusqu'à présent peu exploré dans la littérature scientifique. Nos recherches ont permis d'approfondir cette voie prometteuse et d'apporter des contributions significatives au domaine.

Nous avons démontré que la comparaison intra-patient peut suffire à établir un diagnostic fiable, notamment grâce à l'utilisation de réseaux de neurones auto-supervisés combinés à des algorithmes de détection d'anomalies. Cette approche novatrice offre une nouvelle manière d'aborder les cancers, en les comparant directement entre eux plutôt qu'en analysant uniquement leurs structures individuelles. Les cancers, souvent définis comme des anomalies, peuvent ainsi être identifiés à partir de leurs différences visuelles avec d'autres lésions. L'utilisation de modèles auto-supervisés a également permis une détection des cancers sans nécessiter d'annotations, une avancée majeure. Il reste toutefois aux travaux futurs à étudier l'efficacité de cette méthode pour d'autres pathologies.

Parallèlement, nous avons développé une approche supervisée basée sur les travaux existants, en simplifiant et optimisant les architectures de classification fondées sur des modèles d'attention. Cette simplification a rendu les architectures plus modulables et efficaces, tout en proposant une généralisation aux problématiques similaires. Ces travaux exploitent les dernières avancées en intelligence artificielle, notamment les modèles d'attention et les points d'induction pour l'optimisation des ensembles.

Nos recherches nous ont également conduits à concevoir un modèle novateur pour l'apprentissage des représentations visuelles, exploitant les patchs des Vision Transfor-

mers afin de produire des représentations de haute qualité. L'optimisation simultanée de plusieurs fonctions de perte liées à des problématiques différentes a permis à l'encodeur du modèle de générer des plongements visuels de meilleure qualité. Ces avancées ne se limitent pas à la dermatologie, mais peuvent être appliquées à d'autres domaines, comme l'analyse d'images ou de vidéos. Par ailleurs, nous avons exploré l'interprétabilité des modèles, un enjeu crucial en médecine. Nous avons montré que les techniques de génération d'images conditionnelles peuvent non seulement enrichir la compréhension des classifieurs, mais aussi améliorer leurs performances via l'intégration d'images synthétiques ciblées.

Enfin, la visualisation des vecteurs latents en deux dimensions a permis d'approfondir la compréhension des modèles tout en offrant une représentation intuitive des diagnostics antérieurs. Cette approche facilite l'interprétation et la prise de décision clinique, renforçant ainsi la confiance des praticiens dans ces outils.

En conclusion, cette thèse a posé les bases de nouvelles approches pour le diagnostic du mélanome, en explorant des stratégies innovantes d'apprentissage automatique alliant comparaison intra-patient et représentation visuelle avancée. Nos travaux montrent que ces méthodes peuvent améliorer la précision des diagnostics tout en réduisant la nécessité d'annotations expertes, ouvrant ainsi la voie à une intégration plus fluide de l'intelligence artificielle en médecine. Au-delà de leur application immédiate à la dermatologie, ces avancées peuvent être généralisées à d'autres pathologies et domaines nécessitant une détection fine des anomalies. Nous espérons que ces recherches inspireront de nouvelles avancées, visant à rendre les technologies médicales plus accessibles, performantes et éthiques. Dans cette perspective, les travaux futurs devront affiner et étendre ces méthodes, notamment pour garantir leur adoption en milieu clinique. Plus largement, cette thèse s'inscrit dans une dynamique où l'intelligence artificielle et la médecine convergent vers une complémentarité accrue, au service des praticiens et des patients.

Perspectives

Nous sommes convaincus que nos travaux contribueront à de nouvelles avancées, tant dans le domaine de l'intelligence artificielle que dans celui du diagnostic du mélanome et d'autres pathologies. Les résultats de cette thèse, dont plusieurs ont été publiés dans des conférences et revues scientifiques, visent à nourrir la production de nouvelles recherches au croisement de l'intelligence artificielle et de la médecine.

Notre première approche pourrait trouver une application directe en milieu clinique pour assister les diagnostics dermatologiques. Elle permettrait de fournir un taux d'atypicité pour chaque lésion par patient, tout en ouvrant des possibilités telles que la visualisation et le classement des lésions par ordre d'atypicité ou encore l'affichage de lésions similaires (ayant des représentations latentes proches). En outre, cette méthodologie peut s'adapter à des problématiques d'autres domaines, comme la détection d'anomalies dans le diagnostic d'autres maladies, la fraude financière, ou encore les défauts dans les secteurs industriels, agroalimentaires et logistiques. Des pistes d'amélioration incluent l'intégration de nouveaux modèles d'apprentissage des représentations visuelles, apparus depuis la rédaction de nos travaux, ainsi que l'exploration d'algorithmes complémentaires, tels que les mélanges gaussiens, les forêts d'isolation, DBSCAN ou les facteurs d'éloignement locaux.

Nos travaux supervisés offrent des applications cliniques comparables, mais permettent en plus des prédictions spécifiques entre catégories, comme le mélanome et le naevus. Avec la démocratisation des techniques d'acquisition en 3D, nous sommes convaincus que cette approche, et plus largement les méthodes basées sur la comparaison des lésions, amélioreront la qualité des diagnostics dermatologiques. Les contributions de cette thèse, accessibles en ligne et ouvertes à la communauté scientifique, pourront être réutilisées pour développer de nouveaux outils et approfondir les concepts étudiés.

En ce qui concerne l'apprentissage des représentations visuelles, plusieurs perspectives s'ouvrent pour adapter ce modèle à des données vidéo et évaluer ses performances sur des bases de données variées. Une analyse approfondie des performances en fonction du nombre de paramètres du modèle pourrait également être envisagée, tout comme l'ajout de nouvelles branches dotées de fonctions de perte complémentaires. Une version enrichie du modèle pourrait permettre une modélisation encore plus précise des représentations apprises. Cependant, les défis liés à l'optimisation simultanée de plusieurs fonctions de perte et à la charge computationnelle élevée de ces architectures représentent des problématiques intéressantes pour des travaux futurs.

Enfin, nos travaux portant sur la compréhension des erreurs des réseaux de neurones pourraient renforcer la confiance des cliniciens dans ces outils d'aide au diag-

nostic. Une présentation claire et explicite des erreurs aux experts médicaux pourrait leur permettre de mieux comprendre les forces et les limites des modèles prédictifs, facilitant leur intégration en pratique clinique.

De manière générale, les connaissances acquises au cours de cette thèse fourniront aux chercheurs des pistes pour développer de nouvelles méthodes, améliorer les performances des modèles et accroître l'utilité des outils d'aide au diagnostic, contribuant ainsi à une médecine plus performante et accessible à tous.

Bibliographie

[1] « Cancer Facts & Figures 2024 », *Atlanta*, 2024 (cf. p. 20).

[2] B. STAFF, « Medical gallery of Blausen Medical 2014 », *WikiJournal of Medicine*, t. 1, n° 2, 2014, ISSN : 2002-4436. DOI : 10.15347/wjm/2014.010 (cf. p. 25).

[3] V. ROTEMBERG, N. KURTANSKY, B. BETZ-STABLEIN et al., « A patient-centric dataset of images and metadata for identifying melanomas using clinical context », *Sci Data*, t. 8, p. 34, 2021. DOI : 10.1038/s41597-021-00815-z. adresse : https://doi.org/10.1038/s41597-021-00815-z (cf. p. 28, 36, 37).

[4] J. FERLAY, M. COLOMBET, I. SOERJOMATARAM et al., « Cancer statistics for the year 2020 : An overview », *International journal of cancer*, t. 149, n° 4, p. 778-789, 2021 (cf. p. 29).

[5] A.-M. FORSEA, « Melanoma epidemiology and early detection in Europe : diversity and disparities », *Dermatology practical & conceptual*, t. 10, n° 3, 2020 (cf. p. 29).

[6] H. SUNG, J. FERLAY, R. L. SIEGEL et al., « Global cancer statistics 2020 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries », *CA : a cancer journal for clinicians*, t. 71, n° 3, p. 209-249, 2021 (cf. p. 29).

[7] V. BATAILLE et E. de VRIES, « Melanoma—Part 1 : epidemiology, risk factors, and prevention », *BMJ*, t. 337, 2008, ISSN : 0959-8138. DOI : 10.1136/bmj.a2249. eprint : https://www.bmj.com/content. adresse : https://www.bmj.com/content/337/bmj.a2249 (cf. p. 30).

[8] M. ROSSI, C. PELLEGRINI, L. CARDELLI et al., « Familial Melanoma : Diagnostic and Management Implications », *Dermatology Practical & Conceptual*, t. 9, n° 1, p. 10-16, 2019, PMC6368081, ISSN : 2160-9381. DOI : 10.5826/dpc.0901a03. adresse : https://doi.org/10.5826/dpc.0901a03 (cf. p. 30).

[9] F. LOPES, M. SLEIMAN, K. SEBASTIAN et al., « UV Exposure and the Risk of Cutaneous Melanoma in Skin of Color : A Systematic Review », *JAMA Dermatology*, t. 157, n° 2, p. 213-219, fév. 2021. DOI : 10.1001/jamadermatol.2020.4616 (cf. p. 30).

[10] K. SAGINALA, A. BARSOUK, J. S. ALURU et al., « Epidemiology of Melanoma », *Medical Sciences*, t. 9, n° 4, 2021, ISSN : 2076-3271. DOI : 10.3390/medsci9040063. adresse : https://www.mdpi.com/2076-3271/9/4/63 (cf. p. 30).

[11] H. A. HAENSSLE, C. FINK, R. SCHNEIDERBAUER et al., « Man against machine : diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists », *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, t. 29, n° 8, p. 1836-1842, août 2018, Comparative Study, Journal Article, Observational Study, Validation Study. DOI : 10.1093/annonc/mdy166. adresse : https://www.drks.de/drks_web/ (cf. p. 33).

[12] F. NACHBAR, W. STOLZ, T. MERKLE et al., « The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions », eng, *Journal of the American Academy of Dermatology*, t. 30, n° 4, p. 551-559, 1994, Research Support, Non-U.S. Gov't, ISSN : 0190-9622. DOI : 10.1016/s0190-9622(94)70061-3 (cf. p. 33).

[13] J. J. GROB et J. J. BONERANDI, « The 'Ugly Duckling' Sign : Identification of the Common Characteristics of Nevi in an Individual as a Basis for Melanoma Screening », *Arch Dermatol*, t. 134, n° 1, p. 103-104, jan. 1998. DOI : 10.1001/archderm.134.1.103-a (cf. p. 35).

[14] A. REES et B. FISHER, « A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions », in *Color Medical Image Analysis*, sér. Lecture Notes in Computational Vision and Biomechanics, M. E. CELEBI et G. SCHAEFER, éd., t. 6, Springer, 2013 (cf. p. 36, 37).

[15] D. GUTMAN, N. C. F. CODELLA, E. CELEBI et al., « Skin Lesion Analysis toward Melanoma Detection : A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC) », *arXiv*, 2016. eprint : 1605.01397 (cf. p. 36, 37).

[16] N. CODELLA, D. GUTMAN, M. CELEBI et al., « Skin Lesion Analysis Toward Melanoma Detection : A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC) », *arXiv*, 2017. eprint : 1710.05006 (cs.CV) (cf. p. 36, 37).

[17] N. CODELLA, V. ROTEMBERG, P. TSCHANDL et al., « Skin Lesion Analysis Toward Melanoma Detection 2018 : A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) », *arXiv*, 2018. adresse : https://arxiv.org/abs/1902.03368 (cf. p. 36, 37).

[18] P. TSCHANDL, C. ROSENDAHL et H. KITTLER, « The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions », *Sci. Data*, t. 5, p. 180 161, 2018. DOI : 10.1038/sdata.2018.161 (cf. p. 36, 37).

[19] P. TSCHANDL, C. ROSENDAHL et H. KITTLER, « The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions », *Sci. Data*, t. 5, p. 180 161, 2018. DOI : 10.1038/sdata.2018.161 (cf. p. 36, 37).

[20] N. C. F. CODELLA, D. GUTMAN, M. E. CELEBI et al., « Skin Lesion Analysis Toward Melanoma Detection : A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC) », *arXiv*, 2017. eprint : 1710.05006 (cf. p. 36, 37).

[21] M. COMBALIA, N. C. F. CODELLA, V. ROTEMBERG et al., « BCN20000 : Dermoscopic Lesions in the Wild », *arXiv*, 2019. eprint : 1908.02288 (cf. p. 36, 37).

[22] T. J. BRINKER et al., « Comparing artificial intelligence algorithms to 157 German dermatologists : The melanoma classification benchmark », *European Journal of Cancer*, t. 111, p. 30-37, 2019 (cf. p. 37).

[23] T. MENDONÇA, P. M. FERREIRA, J. MARQUES et al., « PH² - A dermoscopic image database for research and benchmarking », in *Proceedings of the 35th International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, juill. 2013 (cf. p. 37).

[24] W. V. STOECKER, W. W. LI et R. H. MOSS, « Automatic detection of asymmetry in skin tumors », *Computerized Medical Imaging and Graphics*, t. 16, n° 3, p. 191-197, 1992, Digital Imaging in Dermatology, ISSN : 0895-6111. DOI : https://doi.org/10.1016/0895-6111(92)90073-I. adresse : https://www.sciencedirect.com/science/article/pii/089561119290073I (cf. p. 41).

[25] R. G. WHITE, D. A. PEREDNIA et R. A. SCHOWENGERDT, « Automated feature detection in digital images of skin », *Computer Methods and Programs in Biomedicine*, t. 34, n° 1, p. 41-60, 1991, ISSN : 0169-2607. DOI : https://doi.org/10.1016/0169-2607(91)90081-4. adresse : https://www.sciencedirect.com/science/article/pii/0169260791900814 (cf. p. 41).

[26] H. ALQURAN, I. A. QASMIEH, A. M. ALQUDAH et al., « The melanoma skin cancer detection and classification using support vector machine », in *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2017, p. 1-5. DOI : 10.1109/AEECT.2017.8257738 (cf. p. 42).

[27] S. MUSTAFA et A. KIMURA, « A SVM-based diagnosis of melanoma using only useful image features », in *2018 International Workshop on Advanced Image Technology (IWAIT)*, 2018, p. 1-4. DOI : 10.1109/IWAIT.2018.8369646 (cf. p. 42).

[28] C. R. DHIVYAA, K. SANGEETHA, M. BALAMURUGAN et al., « Skin lesion classification using decision trees and random forest algorithms », *Journal of Ambient Intelligence and Humanized Computing*, nov. 2020, ISSN : 1868-5145. DOI : 10.1007/s12652-020-02675-8. adresse : https://doi.org/10.1007/s12652-020-02675-8 (cf. p. 42).

[29] A. MURUGAN, S. A. H. NAIR et K. P. S. KUMAR, « Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers », *Journal of Medical Systems*, t. 43, n° 8, p. 269, juill. 2019, ISSN : 1573-689X. DOI : 10.1007/s10916-019-1400-8. adresse : https://doi.org/10.1007/s10916-019-1400-8 (cf. p. 42).

[30] J. DENG, W. DONG, R. SOCHER et al., « ImageNet : A large-scale hierarchical image database », in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, p. 248-255. DOI : 10.1109/CVPR.2009.5206848 (cf. p. 44).

[31] A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON, « ImageNet Classification with Deep Convolutional Neural Networks », in *Advances in Neural Information Processing Systems*, F. PEREIRA, C. BURGES, L. BOTTOU et al., éd., t. 25, Curran Associates, Inc., 2012. adresse : https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (cf. p. 44).

[32] G. VENKATASUBRAMANIAN, S. KAR, A. SINGH et al., *Towards A Measure Of General Machine Intelligence*, sept. 2021 (cf. p. 44).

[33] S. UMBAUGH, R. MOSS et W. STOECKER, « Automatic color segmentation of images with application to detection of variegated coloring in skin tumors », *IEEE Engineering in Medicine and Biology Magazine*, t. 8, n° 4, p. 43-50, 1989. DOI : 10.1109/51.45955 (cf. p. 45).

[34] J. GOLSTON, R. MOSS et W. STOECKER, « Boundary detection in skin tumor images : An overall approach and a radial search algorithm », *Pattern Recognition*, t. 23, n° 11, p. 1235-1247, 1990, ISSN : 0031-3203. DOI : https://doi.org/10.1016/0031-3203(90)90119-6. adresse : https://www.sciencedirect.com/science/article/pii/0031320390901196 (cf. p. 45).

[35] S. E. ARMAS, M. E. WARKENTIN et O. A. ARMAS, « Expert systems in laboratory medicine and pathology », *Artificial Intelligence in Medicine*, t. 1, n° 2, p. 79-85, 1989, ISSN : 0933-3657. DOI : https://doi.org/10.1016/0933-3657(89)90019-5. adresse : https://www.sciencedirect.com/science/article/pii/0933365789900195 (cf. p. 45).

[36] K. HE, X. ZHANG, S. REN et al., « Deep Residual Learning for Image Recognition », in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cf. p. 45, 57, 65).

[37] K. SIMONYAN et A. ZISSERMAN, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015. arXiv : 1409.1556 [cs.CV]. adresse : https://arxiv.org/abs/1409.1556 (cf. p. 45).

[38] M. TAN et Q. LE, « EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks », in *Proceedings of the 36th International Conference on Machine Learning*, K. CHAUDHURI et R. SALAKHUTDINOV, éd., sér. Proceedings of Machine Learning Research, t. 97, PMLR, juin 2019, p. 6105-6114. adresse : https://proceedings.mlr.press/v97/tan19a.html (cf. p. 45).

[39] A. C. FOAHOM GOUABOU, J.-L. DAMOISEAUX, J. MONNIER et al., « Ensemble Method of Convolutional Neural Networks with Directed Acyclic Graph Using Dermoscopic Images : Melanoma Detection Application », *Sensors*, t. 21, n° 12, 2021, ISSN : 1424-8220. DOI : 10.3390/s21123999. adresse : https://www.mdpi.com/1424-8220/21/12/3999 (cf. p. 45).

[40] H. BHATT, V. SHAH, K. SHAH et al., « State-of-the-art machine learning techniques for melanoma skin cancer detection and classification : a comprehensive review », *Intelligent Medicine*, t. 03, n° 03, p. 180-190, 2023. DOI : 10.1016/j.imed.2022.08.004. eprint : https://mednexus.org/doi/pdf/10.1016/j.imed.2022.08.004. adresse : https://mednexus.org/doi/abs/10.1016/j.imed.2022.08.004 (cf. p. 45).

[41] D. POPESCU, M. EL-KHATIB, H. EL-KHATIB et al., « New Trends in Melanoma Detection Using Neural Networks : A Systematic Review », *Sensors*, t. 22, n° 2, 2022, ISSN : 1424-8220. DOI : 10.3390/s22020496. adresse : https://www.mdpi.com/1424-8220/22/2/496 (cf. p. 45).

[42] Q. ABBAS et M. E. CELEBI, « DermoDeep-A classification of melanoma-nevus skin lesions using multi-feature fusion of visual features and deep neural network », *Multimedia Tools and Applications*, t. 78, n° 16, p. 23 559-23 580, août 2019, ISSN : 1573-7721. DOI : 10.1007/s11042-019-7652-y. adresse : https://doi.org/10.1007/s11042-019-7652-y (cf. p. 45).

[43] J. COLLENNE, J. MONNIER, R. IGUERNAISSI et al., « Fusion between an Algorithm Based on the Characterization of Melanocytic Lesions' Asymmetry with an Ensemble of Convolutional Neural Networks for Melanoma Detection », *Journal of Investigative Dermatology*, t. 144, n° 7, 1600-1607.e2, 2024, ISSN : 0022-202X. DOI : https://doi.org/10.1016/j.jid.2023.09.289. adresse : https://www.sciencedirect.com/science/article/pii/S0022202X24000782 (cf. p. 45, 98).

[44] J. MONNIER, A. C. FOAHOM GOUABOU, M. SERDI et al., « Automated melanoma detection. An algorithm inspired from human intelligence characterizing disordered pattern of melanocytic lesions improving a convolutional neural network », *Journal of the American Academy of Dermatology*, t. 91, n° 2, p. 350-352, 2024, ISSN : 0190-9622. DOI : 10.1016/j.jaad.2024.02.063. adresse : https://doi.org/10.1016/j.jaad.2024.02.063 (cf. p. 45, 130).

[45] Y. FILALI, H. E. KHOUKHI, M. A. SABRI et al., « Efficient fusion of handcrafted and pre-trained CNNs features to classify melanoma skin cancer », *Multimedia Tools and Applications*, t. 79, n° 41, p. 31 219-31 238, nov. 2020, ISSN : 1573-7721. DOI : 10.1007/s11042-020-09637-4. adresse : https://doi.org/10.1007/s11042-020-09637-4 (cf. p. 45).

[46] A. C. FOAHOM GOUABOU, J. COLLENNE, J. MONNIER et al., « Computer Aided Diagnosis of Melanoma Using Deep Neural Networks and Game Theory : Application on Dermoscopic Images of Skin Lesions », *International Journal of Molecular Sciences*, t. 23, n° 22, 2022, ISSN : 1422-0067. DOI : 10.3390/ijms232213838. adresse : https://www.mdpi.com/1422-0067/23/22/13838 (cf. p. 45, 130).

[47] A. ESTEVA, B. KUPREL, R. A. NOVOA et al., « Dermatologist-level classification of skin cancer with deep neural networks », *Nature*, t. 542, n° 7639, p. 115-118, 2017 (cf. p. 45).

[48] H. A. HAENSSLE, C. FINK, R. SCHNEIDERBAUER et al., « Man against machine : diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists », *Annals of Oncology*, t. 29, n° 8, p. 1836-1842, août 2018, Comparative Study, Journal Article, Observational Study, Validation Study, ISSN : 1569-8041. DOI : 10.1093/annonc/mdy166 (cf. p. 45).

[49] T. J. BRINKER, A. HEKLER, J. S. UTIKAL et al., « Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images », *European Journal of Cancer*, t. 113, p. 47-54, 2019 (cf. p. 45).

[50] P. TSCHANDL, C. RINNER, Z. APALLA et al., « Human–computer collaboration for skin cancer recognition », *Nature Medicine*, t. 26, n° 8, p. 1229-1234, 2020 (cf. p. 45).

[51] A. RADFORD, J. W. KIM, C. HALLACY et al., « Learning transferable visual models from natural language supervision », in *International conference on machine learning*, PMLR, 2021, p. 8748-8763 (cf. p. 47).

[52] R. BELLMAN, R. CORPORATION et K. M. R. COLLECTION, *Dynamic Programming* (Rand Corporation research study). Princeton University Press, 1957, ISBN : 9780691079516. adresse : https://books.google.fr/books?id=wdtoPwAACAAJ (cf. p. 50).

[53] R. BELLMAN, *Dynamic Programming* (Dover Books on Computer Science Series). Dover Publications, 2003, ISBN : 9780486428093. adresse : https://books.google.fr/books?id=fyVtp3EMxasC (cf. p. 50).

[54] A. SCOPE, S. W. DUSZA, A. C. HALPERN et al., « The “Ugly Duckling” Sign : Agreement Between Observers », *Archives of Dermatology*, t. 144, n° 1, p. 58-64, jan. 2008, ISSN : 0003-987X. DOI : 10.1001/archdermatol.2007.15. eprint : https://jamanetwork.com/journals/jamadermatology/articlepdf/419304/dst70043_58_64.pdf. adresse : https://doi.org/10.1001/archdermatol.2007.15 (cf. p. 51).

[55] C. GAUDY-MARQUESTE, Y. WAZAEFI, Y. BRUNEU et al., « Ugly Duckling Sign as a Major Factor of Efficiency in Melanoma Detection », *JAMA Dermatology*, t. 153, n° 4, p. 279-284, avr. 2017, ISSN : 2168-6068. DOI : 10.1001/jamadermatol.2016.5500. eprint : https://jamanetwork.com/journals/jamadermatology/articlepdf/2600689/jamadermatology_gaudymarqueste_2017_oi_160073.pdf. adresse : https://doi.org/10.1001/jamadermatol.2016.5500 (cf. p. 51).

[56] M. KRAMER, « Autoassociative neural networks », *Computers Chemical Engineering*, t. 16, n° 4, p. 313-328, 1992, Neutral network applications in chemical engineering, ISSN : 0098-1354. DOI : https://doi.org/10.1016/0098-1354(92)80051-A. adresse : https://www.sciencedirect.com/science/article/pii/009813549280051A (cf. p. 51, 99).

[57] J. GARCIA, S. TANADINI-LANG, N. ANDRATSCHKE et al., « Suspicious Skin Lesion Detection in Wide-Field Body Images using Deep Learning Outlier Detection », in *Annu Int Conf IEEE Eng Med Biol Soc*, juill. 2022, p. 2928-2932. DOI : 10.1109/EMBC48229.2022.9871655 (cf. p. 52).

[58] I. HIGGINS, L. MATTHEY, A. PAL et al., « beta-VAE : Learning Basic Visual Concepts with a Constrained Variational Framework », in *International Conference on Learning Representations*, 2017. adresse : https://openreview.net/forum?id=Sy2fzU9gl (cf. p. 53).

[59] S. PIDHORSKYI, D. A. ADJEROH et G. DORETTO, « Adversarial Latent Autoencoders », in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, [to appear], 2020 (cf. p. 53, 102).

[60] L. R. SOENKSEN, T. KASSIS, S. T. CONOVER et al., « Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images », *Science Translational Medicine*, t. 13, n° 581, eabb3652, 2021. DOI : 10.1126/scitranslmed.abb3652. eprint : https://www.science.org/doi/pdf/10.1126/scitranslmed.abb3652. adresse : https://www.science.org/doi/abs/10.1126/scitranslmed.abb3652 (cf. p. 53, 60).

[61] P. J. CLARK et F. C. EVANS, « Distance to nearest neighbor as a measure of spatial relationships in populations », *Ecology*, t. 35, n° 4, p. 445-453, 1954 (cf. p. 54).

[62] M. ESTER, H.-P. KRIEGEL, J. SANDER et al., « A density-based algorithm for discovering clusters in large spatial databases with noise », in *kdd*, t. 96, 1996, p. 226-231 (cf. p. 54, 55).

[63] B. ZONG, Q. SONG, M. R. MIN et al., « Deep autoencoding gaussian mixture model for unsupervised anomaly detection », in *International conference on learning representations*, 2018 (cf. p. 54).

[64] N. DING, H. MA, H. GAO et al., « Real-time anomaly detection based on long short-Term memory and Gaussian Mixture Model », *Computers & Electrical Engineering*, t. 79, p. 106 458, 2019 (cf. p. 54).

[65] H. VU, D. PHUNG, T. D. NGUYEN et al., « Energy-based models for video anomaly detection », *arXiv preprint arXiv :1708.05211*, 2017 (cf. p. 54).

[66] S. ZHAI, Y. CHENG, W. LU et al., « Deep structured energy based models for anomaly detection », in *International conference on machine learning*, PMLR, 2016, p. 1100-1109 (cf. p. 54).

[67] M. LAFON, E. RAMZI, C. RAMBOUR et al., « Hybrid Energy Based Model in the Feature Space for Out-of-Distribution Detection », *Proceedings of the 40 th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023*, 2023 (cf. p. 54).

[68] Z. YU, V. MAR, A. ERIKSSON et al., « End-to-End Ugly Duckling Sign Detection for Melanoma Identification with Transformers », in sept. 2021, p. 176-184, ISBN : 978-3-030-87233-5. DOI : 10.1007/978-3-030-87234-2_17 (cf. p. 54).

[69] A. VASWANI, N. SHAZEER, N. PARMAR et al., « Attention is All you Need », in *Advances in Neural Information Processing Systems*, t. 30, Curran Associates, Inc., 2017. adresse : https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cf. p. 54).

[70] F. T. LIU, K. M. TING et Z.-H. ZHOU, « Isolation Forest », in *2008 Eighth IEEE International Conference on Data Mining*, 2008, p. 413-422. DOI : 10.1109/ICDM.2008.17 (cf. p. 55).

[71] D. P. KINGMA et M. WELLING, *Auto-Encoding Variational Bayes*, 2022. arXiv : 1312.6114 [stat.ML]. adresse : https://arxiv.org/abs/1312.6114 (cf. p. 55, 100).

[72] X. CHEN et K. HE, « Exploring Simple Siamese Representation Learning », in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2021, p. 15 750-15 758 (cf. p. 57, 58, 66, 77, 80, 83).

[73] M. CARON, I. MISRA, J. MAIRAL et al., « Unsupervised Learning of Visual Features by Contrasting Cluster Assignments », in *Advances in Neural Information Processing Systems*, t. 33, Curran Associates, Inc., 2020, p. 9912-9924. adresse : https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf (cf. p. 58, 66, 77).

[74] T. CHEN, S. KORNBLITH, M. NOROUZI et al., « A Simple Framework for Contrastive Learning of Visual Representations », in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III et A. SINGH, éd., sér. Proceedings of Machine Learning Research, t. 119, PMLR, juill. 2020, p. 1597-1607. adresse : https://proceedings.mlr.press/v119/chen20j.html (cf. p. 58, 66, 76).

[75] J.-B. GRILL, F. STRUB, F. ALTCHÉ et al., « Bootstrap your own latent a new approach to self-supervised learning », in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, sér. NIPS '20, <conf-loc>, <city>Vancouver</city>, <state>BC</state>, <country>Canada</country>, </conf-loc> : Curran Associates Inc., 2020, ISBN : 9781713829546 (cf. p. 58, 66, 77).

[76] L. MCINNES, J. HEALY, N. SAUL et al., « UMAP : Uniform Manifold Approximation and Projection », *Journal of Open Source Software*, t. 3, n° 29, p. 861, 2018. DOI : 10.21105/joss.00861. adresse : https://doi.org/10.21105/joss.00861 (cf. p. 58).

[77] J. LEE, Y. LEE, J. KIM et al., « Set Transformer : A Framework for Attention-based Permutation-Invariant Neural Networks », in *Proceedings of the 36th International Conference on Machine Learning*, K. CHAUDHURI et R. SALAKHUTDINOV, éd., sér. Proceedings of Machine Learning Research, t. 97, PMLR, juin 2019, p. 3744-3753. adresse : https://proceedings.mlr.press/v97/lee19d.html (cf. p. 63, 66).

[78] K. He, H. Fan, Y. Wu et al., « Momentum Contrast for Unsupervised Visual Representation Learning », *arXiv preprint arXiv :1911.05722*, 2019 (cf. p. 66, 75, 76, 83).

[79] C. Doersch, A. Gupta et A. A. Efros, *Unsupervised Visual Representation Learning by Context Prediction*, 2016. arXiv : 1505.05192 [cs.CV]. adresse : https://arxiv.org/abs/1505.05192 (cf. p. 75).

[80] R. Zhang, P. Isola et A. A. Efros, *Colorful Image Colorization*, 2016. arXiv : 1603.08511 [cs.CV]. adresse : https://arxiv.org/abs/1603.08511 (cf. p. 75).

[81] S. Gidaris, P. Singh et N. Komodakis, *Unsupervised Representation Learning by Predicting Image Rotations*, 2018. arXiv : 1803.07728 [cs.CV]. adresse : https://arxiv.org/abs/1803.07728 (cf. p. 75).

[82] A. van den Oord, Y. Li et O. Vinyals, *Representation Learning with Contrastive Predictive Coding*, 2019. arXiv : 1807.03748 [cs.LG]. adresse : https://arxiv.org/abs/1807.03748 (cf. p. 75).

[83] J. Devlin, M.-W. Chang, K. Lee et al., « Bert : Pre-training of deep bidirectional transformers for language understanding », *arXiv preprint arXiv :1810.04805*, 2018 (cf. p. 77).

[84] Y. Liu, M. Ott, N. Goyal et al., « Roberta : A robustly optimized bert pretraining approach », *arXiv preprint arXiv :1907.11692*, 2019 (cf. p. 77).

[85] T. Brown, B. Mann, N. Ryder et al., « Language models are few-shot learners », *Advances in neural information processing systems*, t. 33, p. 1877-1901, 2020 (cf. p. 78).

[86] K. He, X. Chen, S. Xie et al., « Masked Autoencoders Are Scalable Vision Learners », in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, p. 15 979-15 988. DOI : 10.1109/CVPR52688.2022.01553 (cf. p. 78, 79, 82, 85).

[87] A. Jain, P. Abbeel et D. Pathak, « Locally masked convolution for autoregressive models », in *Conference on Uncertainty in Artificial Intelligence*, PMLR, 2020, p. 1358-1367 (cf. p. 78).

[88] R. Köhler, C. Schuler, B. Schölkopf et al., « Mask-specific inpainting with deep neural networks », in *Pattern Recognition : 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, Springer, 2014, p. 523-534 (cf. p. 78).

[89] A. Dosovitskiy, « An image is worth 16x16 words : Transformers for image recognition at scale », *arXiv preprint arXiv :2010.11929*, 2020 (cf. p. 78).

[90] G. Hess, J. Jaxing, E. Svensson et al., « Masked Autoencoder for Self-Supervised Pre-Training on Lidar Point Clouds », in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, jan. 2023, p. 350-359 (cf. p. 78).

[91] C. J. REED, R. GUPTA, S. LI et al., « Scale-MAE : A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning », in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, oct. 2023, p. 4088-4099 (cf. p. 78).

[92] X. CHEN, S. XIE et K. HE, « An Empirical Study of Training Self-Supervised Vision Transformers », in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, oct. 2021, p. 9640-9649 (cf. p. 78).

[93] Y. CHEN, X. SHEN, Y. LIU et al., « Jigsaw-ViT : Learning jigsaw puzzles in vision transformer », *Pattern Recognition Letters*, t. 166, p. 53-60, 2023, ISSN : 0167-8655. DOI : https://doi.org/10.1016/j.patrec.2022.12.023. adresse : https://www.sciencedirect.com/science/article/pii/S0167865522003920 (cf. p. 79).

[94] A. GUPTA, J. WU, J. DENG et al., « Siamese Masked Autoencoders », in *Advances in Neural Information Processing Systems*, A. OH, T. NAUMANN, A. GLOBERSON et al., éd., t. 36, Curran Associates, Inc., 2023, p. 40 676-40 693. adresse : https://proceedings.neurips.cc/paper_files/paper/2023/file/7ffb9f1b57628932518505b5Paper-Conference.pdf (cf. p. 80).

[95] Z. HUANG, X. JIN, C. LU et al., « Contrastive masked autoencoders are stronger vision learners », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023 (cf. p. 80).

[96] A. EYMAËL, R. VANDEGHEN, A. CIOPPA et al., *Efficient Image Pre-Training with Siamese Cropped Masked Autoencoders*, 2024. arXiv : 2403.17823 [cs.CV] (cf. p. 80).

[97] G. GRIFFIN, A. HOLUB et P. PERONA, *Caltech 256*, avr. 2022. DOI : 10.22002/D1.20087 (cf. p. 84).

[98] A. DOSOVITSKIY, L. BEYER, A. KOLESNIKOV et al., « An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale », *ArXiv*, t. abs/2010.11929, 2020. adresse : https://api.semanticscholar.org/CorpusID:225039882 (cf. p. 85, 86).

[99] H. TOUVRON, M. CORD, M. DOUZE et al., « Training data-efficient image transformers amp ; distillation through attention », in *International Conference on Machine Learning*, t. 139, juill. 2021, p. 10 347-10 357 (cf. p. 86).

[100] M. D. ZEILER et R. FERGUS, « Visualizing and Understanding Convolutional Networks », in *Computer Vision – ECCV 2014*, D. FLEET, T. PAJDLA, B. SCHIELE et al., éd., Cham : Springer International Publishing, 2014, p. 818-833, ISBN : 978-3-319-10590-1 (cf. p. 95).

[101] R. R. SELVARAJU, M. COGSWELL, A. DAS et al., « Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization », in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, p. 618-626. DOI : 10.1109/ICCV.2017.74 (cf. p. 96).

[102] A. CHATTOPADHAY, A. SARKAR, P. HOWLADER et al., « Grad-CAM++ : Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks », in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, p. 839-847. DOI : 10.1109/WACV.2018.00097 (cf. p. 96).

[103] H. WANG, Z. WANG, M. DU et al., « Score-CAM : Score-Weighted Visual Explanations for Convolutional Neural Networks », in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, p. 111-119. DOI : 10.1109/CVPRW50498.2020.00020 (cf. p. 96).

[104] K. MRIDHA, M. M. UDDIN, J. SHIN et al., « An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System », *IEEE Access*, t. 11, p. 41 003-41 018, 2023. DOI : 10.1109/ACCESS.2023.3269694 (cf. p. 97).

[105] J. A. MARMOLEJO-SAUCEDO et U. KOSE, « Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis », *Mobile Networks and Applications*, t. 27, n° 4, p. 879-893, août 2022, ISSN : 1572-8153. DOI : 10.1007/s11036-022-02021-6. adresse : https://doi.org/10.1007/s11036-022-02021-6 (cf. p. 97).

[106] H. PANWAR, P. GUPTA, M. K. SIDDIQUI et al., « A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images », *Chaos, Solitons Fractals*, t. 140, p. 110 190, 2020, ISSN : 0960-0779. DOI : https://doi.org/10.1016/j.chaos.2020.110190. adresse : https://www.sciencedirect.com/science/article/pii/S0960077920305865 (cf. p. 97).

[107] M. D. ZEILER et R. FERGUS, « Visualizing and Understanding Convolutional Networks », in *Computer Vision – ECCV 2014*, D. FLEET, T. PAJDLA, B. SCHIELE et al., éd., Cham : Springer International Publishing, 2014, p. 818-833, ISBN : 978-3-319-10590-1 (cf. p. 97, 98).

[108] M. T. RIBEIRO, S. SINGH et C. GUESTRIN, « "Why Should I Trust You?" : Explaining the Predictions of Any Classifier », in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, sér. KDD '16, San Francisco, California, USA : Association for Computing Machinery, 2016, p. 1135-1144, ISBN : 9781450342322. DOI : 10.1145/2939672.2939778. adresse : https://doi.org/10.1145/2939672.2939778 (cf. p. 98).

[109] N. NIGAR, M. UMAR, M. K. SHAHZAD et al., « A deep learning approach based on explainable artificial intelligence for skin lesion classification », *IEEE Access*, t. 10, p. 113 715-113 725, 2022 (cf. p. 98).

[110] A. XIANG et F. WANG, « Towards interpretable skin lesion classification with deep learning models », in *AMIA annual symposium proceedings*, American Medical Informatics Association, t. 2019, 2019, p. 1246 (cf. p. 98).

[111] S. M. LUNDBERG et S.-I. LEE, « A Unified Approach to Interpreting Model Predictions », in *Advances in Neural Information Processing Systems 30*, I. GUYON, U. V. LUXBURG, S. BENGIO et al., éd., Curran Associates, Inc., 2017, p. 4765-4774. adresse : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (cf. p. 98).

[112] S. WANG, Y. YIN, D. WANG et al., « Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis », *IEEE Transactions on Cybernetics*, t. 52, n° 12, p. 12 623-12 637, 2022. DOI : 10.1109/TCYB.2021.3069920 (cf. p. 98).

[113] A. NGUYEN, J. YOSINSKI et J. CLUNE, *Multifaceted Feature Visualization : Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks*, 2016. arXiv : 1602.03616 [cs.NE]. adresse : https://arxiv.org/abs/1602.03616 (cf. p. 99).

[114] I. J. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA et al., *Generative Adversarial Networks*, 2014. arXiv : 1406.2661 [stat.ML] (cf. p. 99).

[115] R. ROMBACH, A. BLATTMANN, D. LORENZ et al., « High-Resolution Image Synthesis With Latent Diffusion Models », in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2022, p. 10 684-10 695 (cf. p. 99, 101, 104).

[116] A. Q. NICHOL et P. DHARIWAL, « Improved Denoising Diffusion Probabilistic Models », in *Proceedings of the 38th International Conference on Machine Learning*, M. MEILA et T. ZHANG, éd., sér. Proceedings of Machine Learning Research, t. 139, PMLR, juill. 2021, p. 8162-8171. adresse : https://proceedings.mlr.press/v139/nichol21a.html (cf. p. 99).

[117] I. J. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA et al., *Generative Adversarial Networks*, 2014. arXiv : 1406.2661 [stat.ML] (cf. p. 100).

[118] M. MIRZA et S. OSINDERO, *Conditional Generative Adversarial Nets*, 2014. arXiv : 1411.1784 [cs.LG]. adresse : https://arxiv.org/abs/1411.1784 (cf. p. 101).

[119] J.-Y. ZHU, T. PARK, P. ISOLA et al., « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks », *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 2242-2251, 2017. adresse : https://api.semanticscholar.org/CorpusID:206770979 (cf. p. 101).

[120] J. SOHL-DICKSTEIN, E. A. WEISS, N. MAHESWARANATHAN et al., « Deep unsupervised learning using nonequilibrium thermodynamics », in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, sér. ICML'15, Lille, France : JMLR.org, 2015, p. 2256-2265 (cf. p. 101).

[121] J. HO, A. JAIN et P. ABBEEL, *Denoising Diffusion Probabilistic Models*, 2020. arXiv : 2006.11239 [cs.LG]. adresse : https://arxiv.org/abs/2006.11239 (cf. p. 101).

[122] T. KARRAS, S. LAINE et T. AILA, *A Style-Based Generator Architecture for Generative Adversarial Networks*, 2019. arXiv : 1812.04948 [cs.NE]. adresse : https://arxiv.org/abs/1812.04948 (cf. p. 102).

[123] T. KARRAS, S. LAINE, M. AITTALA et al., *Analyzing and Improving the Image Quality of StyleGAN*, 2020. arXiv : 1912.04958 [cs.CV]. adresse : https://arxiv.org/abs/1912.04958 (cf. p. 102).

[124] T. KARRAS, M. AITTALA, S. LAINE et al., « Alias-Free Generative Adversarial Networks », in *Proc. NeurIPS*, 2021 (cf. p. 102).

[125] H. RASHID, M. A. TANVEER et H. AQEEL KHAN, « Skin Lesion Classification Using GAN based Data Augmentation », in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, p. 916-919. DOI : 10.1109/EMBC.2019.8857905 (cf. p. 103).

[126] Z. QIN, Z. LIU, P. ZHU et al., « A GAN-based image synthesis method for skin lesion classification », *Computer Methods and Programs in Biomedicine*, t. 195, p. 105 568, 2020, ISSN : 0169-2607. DOI : https://doi.org/10.1016/j.cmpb.2020.105568. adresse : https://www.sciencedirect.com/science/article/pii/S0169260720302418 (cf. p. 103).

[127] C. SCHUHMANN, R. VENCU, R. BEAUMONT et al., « LAION-400M : Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs », in *Proceedings of Neurips Data-Centric AI Workshop*, 2021 (cf. p. 104).

[128] A. MIKOŁAJCZYK, S. MAJCHROWSKA et S. CARRASCO LIMEROS, « The (de)biasing Effect of GAN-Based Augmentation Methods on Skin Lesion Images », in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. WANG, Q. DOU, P. T. FLETCHER et al., éd., Cham : Springer Nature Switzerland, 2022, p. 437-447, ISBN : 978-3-031-16452-1 (cf. p. 105).

[129] L. MCINNES, J. HEALY, N. SAUL et al., « UMAP : Uniform Manifold Approximation and Projection », *Journal of Open Source Software*, t. 3, n° 29, p. 861, 2018. DOI : 10.21105/joss.00861. adresse : https://doi.org/10.21105/joss.00861 (cf. p. 109).

[130] J. COLLENNE, J. MONNIER, R. IGUERNAISSI et al., « Fusion between an Algorithm Based on the Characterization of Melanocytic Lesions' Asymmetry with an Ensemble of Convolutional Neural Networks for Melanoma Detection », *Journal of Investigative Dermatology*, 2024, ISSN : 0022-202X. DOI : https://doi.org/10.1016/j.jid.2023.09.289. adresse : https://www.sciencedirect.com/science/article/pii/S0022202X24000782 (cf. p. 130).

ANNEXES

A. Travaux sur l'asymétrie des lésions cutanées

Les travaux de la présente thèse s'appuient également sur la base de précédents travaux que nous avons réalisé au sein du Laboratoire Informatique et Systèmes durant un stage de recherche, avant de débuter la thèse. Ces travaux ont permis la publication de l'article "*Fusion between an Algorithm Based on the Characterization of Melanocytic Lesions' Asymmetry with an Ensemble of Convolutional Neural Networks for Melanoma Detection*" [130] dans la revue *Journal Of Investigative Dermatology*. L'idée générale est que nous avons développé un système expert mesurant l'asymétrie au sein des lésions. La sortie de ce système, qui est une mesure de l'asymétrie, et ensuite concatenée avec les représentations apprises des images par un CNN, puis donnée en entrée à un classifieur. De cette manière, le modèle final comporte une mesure de l'asymétrie ainsi qu'un diagnostic, permettant d'analyser les contributions de l'asymétrie au sain des lésions malignes. Ces travaux sont en lien avec la thèse, fusionnant les systèmes experts et les réseaux de neurones, et présentant une amélioration des résultats et une interprétabilité des modèles accrue.

B. Travaux sur la théorie des jeux appliquée à la détection de maladies cutanées

Nous avons également eu la chance de collaborer avec d'autres chercheurs durant cette thèse, notamment avec Arthur Cartel Foahom Gouabou, titulaire d'un doctorat à Aix-Marseille Université, sur l'application de la théorie des jeux dans la détection de maladies cutanées. Cet article présente une nouvelle architecture basée sous la forme d'un arbre de réseaux de neurones, tous spécialisés dans certains types de maladies. En divisant le problème initial du diagnostic sur plusieurs maladies, l'algorithme permet de meilleures performances sur certains types de maladies difficiles à diagnostiquer. Ces travaux ont permi la publication de l'article "*Computer Aided Diagnosis of Melanoma Using Deep Neural Networks and Game Theory : Application on Dermoscopic Images of Skin Lesions*" [46] dans la revue *International Journal of Molecular Sciences*.

C. Travaux sur la notion de désordre en dermatologie

Enfin, nous avons également travaillé avec Jilliana Monnier sur l'analyse du désordre au sein des lésions cutanées pour la détection du mélanome. Ces travaux sont publiés au journal "*Journal of the American Academy of Dermatology*" dans l'article "*Automated melanoma detection. An algorithm inspired from human intelligence characterizing disordered pattern of melanocytic lesions improving a convolutional neural network*" [44].