E4FI - Mini projet

Machine Learning et Cancer du poumon

Étude des Techniques de Détection

Charles BATCHAEV
Nathalia PEREZ RAMIREZ
Steeve RABEHANTA
Élodie PAN
Inès DAKKAK

Table des matières

1. Introduction

Le cancer du poumon demeure un problème de santé publique mondial majeur, étant la principale cause de décès par cancer. Selon le Centre International de Recherche sur le Cancer (CIRC), également connu sous le nom d'International Agency for Research on Cancer (IARC), on estime qu'en 2020, il a été responsable de 1,8 million de décès, soit 18% de l'ensemble des décès liés au cancer. Cette maladie se décline principalement sous deux formes : le Carcinome Non à Petites Cellules (CPNPC) ou NSCLC en anglais, et le Carcinome à Petites Cellules (CPC), SPC en anglais. Le CPNPC est plus répandu et évolue lentement, tandis que le CPC est moins courant mais a tendance à se développer rapidement. (1)

Le cancer du poumon, en raison de son incidence et de sa mortalité élevées, représente un défi de taille pour le domaine de la recherche médicale. Cependant, la détection précoce de cette maladie peut considérablement améliorer les taux de survie. Selon les données scientifiques actuelles, au moins 40 % de tous les cas de cancer pourraient être évités grâce à des mesures de prévention primaire efficaces, et la mortalité supplémentaire peut être réduite grâce à la détection précoce des tumeurs. (2)

De nos jours, les méthodes dites traditionnelles de dépistage des cancers comme la radiographie sont complétées par des technologies comme les CT-scan (computed-tomography). Avec l'évolution fulgurante de l'intelligence artificielle et des CNN (convolutional neural networks), les dépistages de cancer sont davantage facilités. (3)

Les CT-scans sont des scans à haute résolution permettant de dépister des nodules pulmonaires suspects de manière précoce. A l'aide de l'imagerie, les médecins vont chercher des tissus internes qui contiennent potentiellement des anomalies. Des algorithmes comme les CNN, correspondant au traitement de données d'imagerie et de vision par ordinateur, permettent d'identifier le type des nodules pulmonaires ce qui va permettre aux médecins d'effectuer un dépistage rapide et pertinent du cancer. Par la suite, des modèles de machine learning et deep learning ont été mis en place pour améliorer la détection de cancer afin d'améliorer la précision des résultats et diminuer le plus possible les taux de faux positifs. (3)

Ainsi, les avancées dans le monde médical permettent d'anticiper le stade d'évolution du cancer des poumons grâce à l'imagerie, permettant le traitement rapide des patients atteints. Cependant, de nombreuses controverses persistent quant à la place des IA dans ce domaine. De nombreux membres du corps médical se posent des questions éthiques dans son utilisation, particulièrement quant à la confidentialité des informations des patients et à la fiabilité des IA. (3)

2. État de l’art - Techniques d’analyse

2.1. Prétraitement des données

Dans le domaine de la prédiction de l'étape pathologique dans le cancer du poumon non à petites cellules (CPNPC/NSCLC en anglais), la phase de prétraitement des données revêt une importance cruciale. Une étape initiale essentielle consiste à vérifier si la distribution des classes d'origine de la cohorte CPNPC est équilibrée. En cas de déséquilibre, il est courant de recourir à une technique d'oversampling, telle que l'algorithme SMOTE (Synthetic Minority Over-sampling Technique). Cette approche vise à rétablir l'équilibre de la distribution des classes en générant un nouvel ensemble de données synthétiques. Les jeux de données ainsi créés sont ensuite divisés en deux ensembles distincts : un ensemble d'entraînement, qui servira à former le modèle, et un ensemble de tests, qui sera utilisé pour évaluer la performance du modèle. Cette phase de prétraitement sert de base essentielle pour le développement ultérieur de l'algorithme de Machine Learning dédié à la prédiction de l'étape pathologique du CPNPC. (4)

2.2. Segmentation

La segmentation par CT-scans joue un rôle important dans le diagnostic du cancer du poumon. Cette étape permet de détecter automatiquement les motifs suspects et classe les anomalies présentes dans l'échantillon donnée. Il en existe plusieurs types. Tout d'abord il y a la segmentation sémantique, qui est la plus utilisée. Elle consiste à segmenter l'image en différents groupes de pixels, ce qui permet de délimiter les contours des tumeurs. Ensuite, il y a les CNN (Convolutional Neural Network), et particulièrement le modèle U-Net, qui apprennent des caractéristiques complexes à partir d'images. Cette méthode permet de délimiter les zones d'intérêts avec précision, c'est pourquoi ce type de segmentation est souvent utilisé dans le milieu médical. Enfin, pour traiter les images tridimensionnelles, il y a l'équivalent des CNN, les 3D-Res2Unet. (5)(6)(7)

2.3. Techniques de classification

Une méthode permettant de détecter le cancer du poumon via des images est le système des Réseaux de Neurones Artificiels, autrement dit ANN. L’ANN est une

méthode dite statistique et se base sur l'imitation du réseau neuronales complexes humains. La méthode reprend les systèmes d'apprentissage de l'homme, avec plusieurs couches neuronales et se base sur deux niveaux : l'entraînement puis la validation. Cela lui permet d'être efficace lors des traitements de données complexes qui ne sont pas forcément linéaires (comme dans notre cas avec la détection du cancer du poumon). (8)

Pour commencer, la méthode s'applique sur des images qui ont été traitées au préalable afin d'optimiser au maximum leur qualité et ainsi pouvoir par la suite les classifier afin de réaliser les prédictions. (9)

Les Réseaux vont ainsi traiter les images : chaque pixel de l'image est converti en forme numérique, ce qui va permettre son analyse via ce nouvel ensemble de données. Dans notre cas, l'image permettant la détection du cancer du poumon, qui peut être une tomodensitométrie de la poitrine par exemple, va être mise dans le réseau. Toutes les couches de ce réseau vont appliquer plusieurs transformations mathématiques. A chaque couche, le réseau va identifier des caractéristiques de plus en plus difficiles à reconnaître. (10)

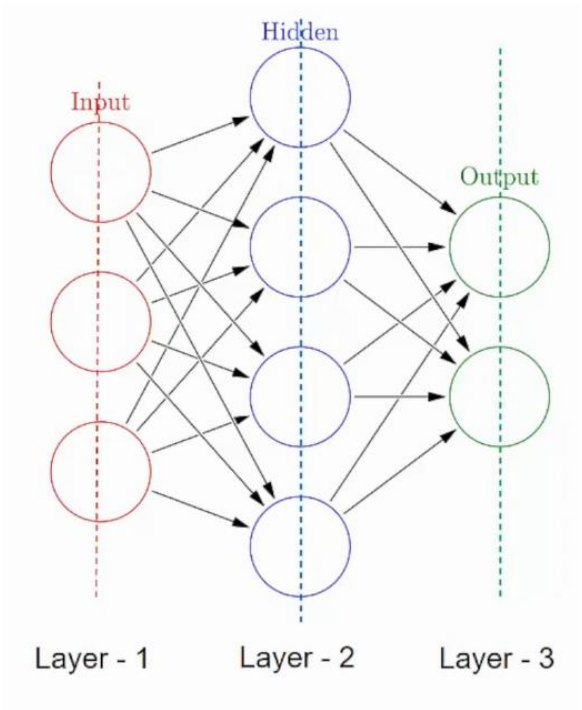

Figure 1. Fonctionnement du réseau représentant les neurones artificiels (10)

Par exemple, les premières couches peuvent détecter des bordures ou des textures de cellule cancéreuse. Les couches suivantes, quant à elles, vont analyser des structures médicales plus complexes comme des nodules ou des masses suspectées comme cancérigènes. Les ANN ont un processus d'apprentissage dit supervisé : le réseau va dans un premier temps être entraîné sur un grand nombre d'images labellisées.

Grâce à cet apprentissage, le réseau pourra ensuite reconnaître des caractéristiques que l'on aura associées au cancer des poumons mais cette fois-ci sur des nouvelles images. (11)

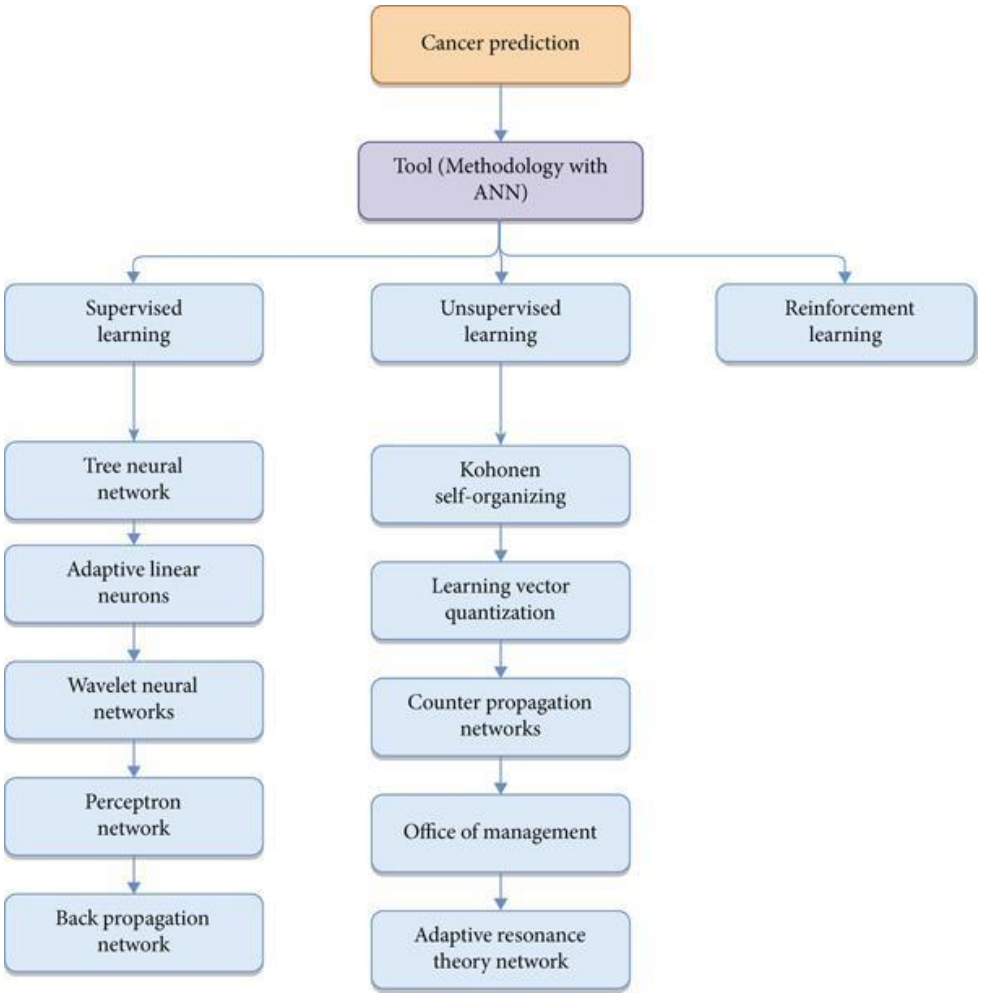

Figure 2. Cancer prediction using ANN tools. (12)

L'avantage de l'ANN est principalement sa capacité rapide à apprendre des caractéristiques pertinentes directement grâce aux données. Cependant, l'ANN peut

avoir des difficultés quand il s'agit d'interpréter des modèles. De plus, pour son apprentissage supervisé, il faut fournir au réseau de très grandes quantités de données pour qu'il puisse fonctionner correctement.

Les modèles ANN permettent d'analyser les données de manière efficace. Avec les avancées technologiques ainsi que celles du machine learning, d'autres techniques ont été développées visant à être plus précis et efficaces concernant l'analyse et le traitement d'image, telles que les Réseaux Neuronaux Convolutifs (CNN).

Les CNN filtrent de la même manière que les ANN, mais de façon plus complexe. En effet, les CNN vont filtrer à l'aide de filtre par convolution et par max pooling pour traiter l'image et réaliser la détection du cancer.

Tout d'abord pour traiter l'image, les CNN vont appliquer la convolution. Ainsi, des filtres vont être appliqués sur l'image pour spécifier les caractéristiques de l'image traitées comme des bords ou encore textures. Tous les filtres, petites sections par petites sections, vont faire un produit scalaire entre les valeurs des pixels et du filtre. Ainsi, une image "carte de caractéristiques" est créée, sur laquelle la convolution va être appliquée. Cela va permettre au réseau d'apprendre des relations complexes. (13)

Par la suite, les CNN vont appliquer le pooling. La carte de caractéristiques va alors conserver uniquement les informations essentielles de l'image. Pour ce faire, le max pooling garde uniquement les valeurs maximales des sections de la carte. Après plusieurs cycles de convolution et de pooling appliqués, les cartes sont écrasées et transmises aux couches connectées, à l'instar des ANN. Ces dernières couches apprennent des relations entre les données ce qui va permettre de faire les prédictions finales (ici pour calculer la probabilité de la présence d'un cancer du poumon). (12)(13)(14)

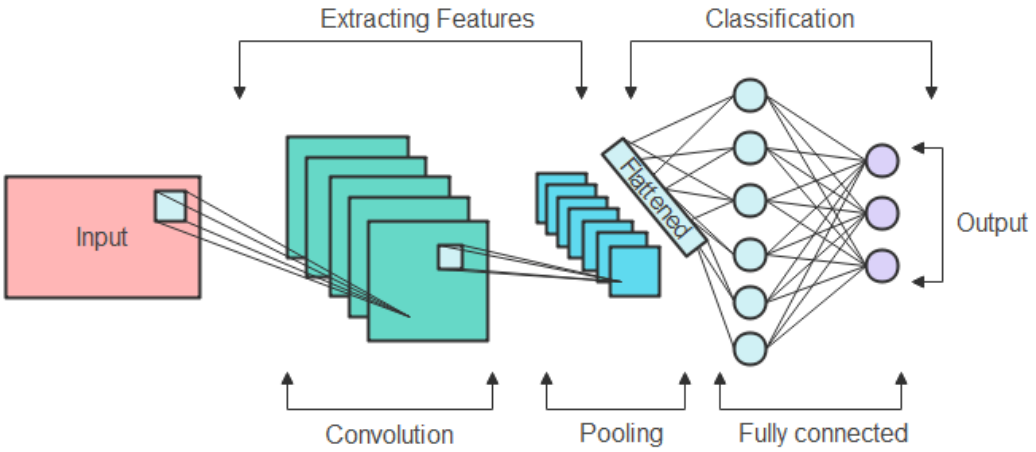

Figure 3. Système CNN. (15)

En parallèle, la méthode des K plus proches voisins (KPPV), ou K-nearest neighbours (KNN) en anglais, émerge comme un algorithme d'apprentissage supervisé non-paramétrique dédié à la résolution de problèmes de classification et de régression. En contraste avec les CNN, le KPPV se distingue par son caractère non-paramétrique, se basant exclusivement sur les données d'entraînement, à l'exception de K, qui doit être préalablement fixé. Dans le cadre de l'apprentissage supervisé, le modèle est formé avec des données étiquetées, ce qui habilite l'algorithme à classifier des données et à prédire des résultats avec précision.

La méthode des K plus proches voisins se base sur une idée très simple : à partir d'un jeu de données, déterminer les plus proches voisins de l'entrée x en calculant la distance de chaque point de donnée à x.

Cet algorithme fait partie de la famille de modèles "d'apprentissage paresseux" car il stocke seulement un jeu de données pour effectuer ses prédictions au lieu de passer par une phase d'entraînement. Ainsi, tous les calculs sont effectués au moment de la classification ou de la prédiction. De plus, l'algorithme a fortement recours à la mémoire pour stocker les données d'entraînement.

En classification, autrement dit avec des variables discrètes, on attribue à x la classe dominante parmi celles des voisins identifiés.

Quant à la régression, avec des variables continues, la valeur prédite est généralement la moyenne, la médiane ou la variance des plus proches voisins.

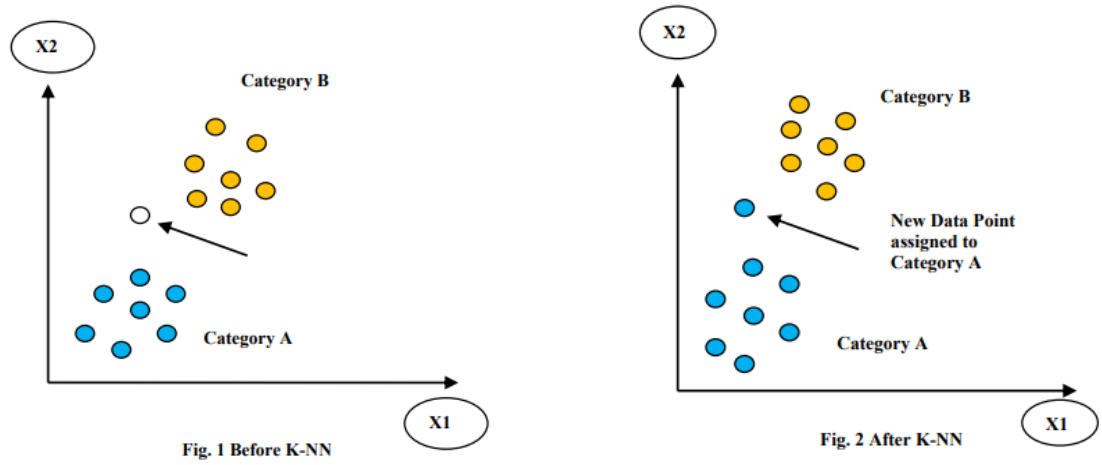

Figure 4. Classification en utilisant la méthode KNN.

Pour calculer la distance, deux méthodes sont principalement utilisées (16) :

- la distance Euclidienne $\sum_{i=1}^{n} (x_i - y_i)^2$
- la distance Manhattan $\sum_{i=1}^{n} |x_i - y_i|$

avec x : le point observé

y : le point de référence

n : la dimension de l'espace

Avantages :

La méthode des K plus proches voisins est très facile à mettre en œuvre car il n'est pas nécessaire de créer un modèle ou de régler des paramètres. Tout dépend du jeu de données fourni. De plus, il s'agit d'un algorithme polyvalent puisqu'il peut être utilisé pour des problèmes de classification ou de régression. (16)(17)

Inconvénients :

Étant donné qu'il repose sur le jeu de données d'apprentissage, plus la taille du jeu de données est importante, plus l'algorithme sera coûteux en temps mais aussi en argent.

De même, le nombre k de voisins est aussi un facteur à prendre en compte car il influe sur les calculs effectués par l'algorithme. (18)

3. Analyse et simulation de la méthode SMOTE

Avec l'état de l'art de la détection du cancer des poumons, nous constatons une variété des techniques utilisant le machine learning. Les ANN incluant les CNN et les approches classiques comme les KNN permettent de comprendre dans un premier temps les bases de la classification des types de cancer sur des images médicales. L'augmentation des données (Data Augmentation), la méthode SMOTE et l'utilisation de poids de classes complètent ces modèles de base pour les rendre plus performants. Désormais, nous allons réaliser une analyse détaillée de la méthode SMOTE, de l'augmentation des données et de l'application des poids de classes. Cette analyse va nous permettre de comprendre en détail les mécanismes de ces méthodes dans la détection du cancer du poumon. Nous allons également reproduire des simulations concrètes afin de mesurer leurs performances. (*Voir Note Book)*

4. Conclusion

En conclusion, les techniques de machine learning comme SMOTE, la Data augmentation ou encore le poids de classes révèlent que l'utilisation de SMOTE se distingue par sa capacité à améliorer la précision des résultats ainsi que le F1-score dans le contexte de la détection du cancer du poumon à partir d'images. Cependant, il faut prendre en compte que l'efficacité de la technique SMOTE dépend largement de la qualité des échantillons synthétiques générés.

Face à cette mission donnée par notre entreprise, les résultats obtenus lors de nos tests et analyses nous offrent une voie prometteuse quant à l'amélioration de nos outils permettant la détection du cancer du poumon de manière efficace et automatisée.

5. Références

1. World Health Organization : WHO & World Health Organization : WHO. (2023, 26 juin). Lung cancer. https://www.who.int/news-room/fact-sheets/detail/lung-cancer

2. Cancer Topics – IARC. (s. d.). https://www.iarc.who.int/cancer-topics/

3. S. Poonkodi, M. Kanchana - *A review on lung carcinoma segmentation and classification using CT image based on deep learning* (2022). *https://www.inderscienceonline.com/doi/epdf/10.1504/IJISTA.2022.125608*

4. Yu, L., Tao, G., Zhu, L. *et al.* Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. *BMC Cancer* 19, 464 (2019). https://doi.org/10.1186/s12885-019-5646-9

5. Çifçi, M. A. (2022). SEGChANET : A novel model for lung cancer segmentation in CT scans. *Applied Bionics and Biomechanics*, *2022*, 1-16. https://doi.org/10.1155/2022/1139587

6. Roy, A., & Todorovic, S. (2016). A multi-scale CNN for affordance segmentation in RGB images. Dans *Lecture Notes in Computer Science* (p. 186-201). https://doi.org/10.1007/978-3-319-46493-0_12

7. Kassel, R. (2023, 9 novembre). *U-NET : le réseau de neurones de computer vision*. Formation Data Science | DataScientest.com. https://datascientest.com/u-net

8. *What are neural networks ?* | *IBM*. (s. d.). https://www.ibm.com/topics/neural-networks

9. Prisciandaro, E., Sedda, G., Cara, A., Diotti, C., & Spaggiari, L. (2023). Artificial Neural Networks in lung Cancer Research : A Narrative review. *Journal of Clinical Medicine*, *12*(3), 880. https://doi.org/10.3390/jcm12030880

10. Singh, H. (2023, 27 juillet). *Deep Learning 101 : Beginners guide to neural network*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/basics-of-neural-network/

11. Nasser, I. M. (2019). ANN for Lung Cancer Detection. *International Journal of Engineering and Information Systems (IJEAIS)*, *3*(3), 17-21. https://philarchive.org/archive/ALIAFL

12. Gangurde, R., Jagota, V., Khan, M. S., Sakthi, V. S., Boppana, U. M., Osei, B., & Kakarla, H. K. (2023). Developing an efficient cancer detection and prediction tool using Convolution neural network integrated with neural pattern recognition. *BioMed Research International*, *2023*, 1-11. https://doi.org/10.1155/2023/6970256

13. Zhu, Z., Wang, S., & Zhang, Y. (2023). A survey of convolutional neural network in breast cancer. *Cmes-computer Modeling in Engineering & Sciences*, *136*(3), 2127-2172. https://doi.org/10.32604/cmes.2023.025484

14. Science, B. O. C., & Science, B. O. C. (2023, 24 mai). *What is the purpose of a feature map in a convolutional neural network | Baeldung on Computer Science*. Baeldung on Computer Science. https://www.baeldung.com/cs/cnn-feature-map

15. *Neural Network Diagram Complete Guide | EDrawMax*. (s. d.). Edrawsoft. https://www.edrawsoft.com/article/neural-network-diagram.html

16. Suyal, M., & Goyal, P. (2022). A review on analysis of K-Nearest Neighbor Classification machine learning algorithms based on supervised learning. *International journal of engineering trends and technology*, *70*(7), 43-48. https://doi.org/10.14445/22315381/ijett-v70i7p205

17. Robert, J. (2023, 9 novembre). *KNN : Découvrez cet algorithme de machine learning*. Formation Data Science | DataScientest.com. https://datascientest.com/knn

18. Turpin, A. (2023, 9 novembre). *Qu'est-ce que le KNN ? Le modèle de machine learning supervisé*. https://www.jedha.co/formation-ia/algorithme-knn-apprentissage-supervise