الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Constantine 1 Frères Mentouri — **الإخوة منتوري1 جامعة قسنطينة**

Faculté des Sciences de la Nature et de la Vie — **كلية علوم الطبيعة والحياة**

Département : Biologie Appliquée — **قسم البيولوجيا التطبيقية**

Mémoire présenté en vue de l'obtention du Diplôme de Master

Domaine : Sciences de la Nature et de la Vie

Filière : Biotechnologies

Spécialité : Bio-Informatique

N° d'ordre :
N° de série :

Thème :

Prédiction du cancer des poumons à partir des données d'expression des gènes basée sur le Deep learning

Présenté par : SAKLOUL Malak **Le : 25/06/2025**

Jury d'évaluation :

Président : Dr. DAAS Mohamed Skander

Directeur de thèse : Pr. BELLIL INES (Univ.Constantine 1 Frères Mentouri)

Examinatrice : Dr. AMINE KHOUDJA ihsein Rokia

Année universitaire 2024 - 2025

REMERCIEMENTS

Je remercie du fond du cœur **ALLAH**, le Tout-Puissant, de m'avoir accompagnée tout au long de ce parcours.

C'est par Sa grâce que j'ai trouvé la volonté, la force et le courage nécessaires pour mener à bien ce mémoire.

Dans les moments de doute, Il m'a offert la patience ; dans les instants d'épreuve, Il m'a accordé la sérénité

Je tiens à exprimer mes profonds remerciements à toutes les personnes qui ont contribué à la réalisation de ce mémoire de fin d'étude.

Je tiens à adresser mes remerciements les plus sincères à mon directeur de mémoire **Pr.BELLIL INES** pour sa patience, sa disponibilité et sa bienveillance tout au long de ce travail.

Je suis pleinement consciente que ma progression n'a pas toujours été régulière, et je lui suis profondément reconnaissante pour sa compréhension et sa flexibilité face à mes retards et mes hésitations.

Sa capacité à m'encadrer avec rigueur, tout en faisant preuve d'une grande humanité, a été un soutien précieux dans les moments de doute.

Merci pour votre confiance, votre indulgence et vos conseils éclairés qui ont grandement contribué à la réalisation de ce mémoire.

Je tiens à exprimer toute ma gratitude à **Dr. ALIOUANE SALAH EDDINE** pour son accompagnement précieux tout au long de ce travail.

Merci pour votre soutien constant, votre rigueur intellectuelle et votre grande maîtrise dans l'orientation de mes recherches.

Votre capacité à cerner l'essentiel, à guider avec clarté et à encourager avec justesse a été un véritable moteur dans l'élaboration de ce mémoire.

Votre implication, votre écoute et votre concentration attentive ont grandement contribué à donner à ce travail sa forme actuelle.

Je remercie également les membres du jury (Dr. DAAS Mohamed Skander et Dr. AMINE KHOUDJA ihsein Rokia) pour le temps qu'ils ont consacré à l'évaluation de ce mémoire, ainsi que pour leurs remarques constructives et leurs suggestions pertinentes qui ont contribué à enrichir et améliorer la qualité de ce travail.

Enfin, je tiens à exprimer toute ma reconnaissance à l'ensemble de l'équipe pédagogique de l'Université de Constantine 1, et plus particulièrement aux professeurs de bio-informatique, dont le soutien, les enseignements et la disponibilité m'ont accompagnée tout au long de mes études.

Dédicaces

À mes parents

À vous, mes parents bien-aimés,

Il n'existe pas de mots assez forts pour décrire toute ma reconnaissance.

Merci pour votre amour inconditionnel, pour votre présence constante, pour vos sacrifices silencieux que j'ai parfois compris bien trop tard.

Vous avez cru en moi même quand je doutais de moi-même. Vous m'avez soutenue, guidée, encouragée, relevée, aimée... sans jamais attendre en retour.

C'est grâce à votre force, votre patience et votre foi en moi que j'ai pu franchir chaque étape de ce chemin parfois difficile.

Vous êtes les fondations de tout ce que je suis aujourd'hui.

Ce travail, ce diplôme, cette réussite… c'est aussi la vôtre.

Merci du fond du cœur. Je vous dois bien plus qu'un remerciement : je vous dois ma réussite, mon courage, et ma sérénité.

À ma sœur Imen

Merci à ma sœur Imen, pour ta présence rassurante, ton écoute, tes encouragements et ton affection. Tu as su m'apporter du réconfort dans les moments difficiles et partager ma joie dans les moments de réussite. Ta bienveillance m'a beaucoup aidée.

À mon frère Aymen

Un grand merci à mon frère Aymen, pour ton soutien discret mais toujours présent. Ta force tranquille et ton exemple m'ont inspirée à persévérer dans les moments de doute. Je suis fierère de t'avoir comme frère.

À mon frère Bahaa

Merci du fond du cœur à mon frère Bahaa, pour ta générosité, ta bonne humeur et ta confiance en moi. Ta présence a toujours été un appui moral précieux, et ton humour a souvent su alléger la pression.

À mes cousines Hanna, Rayen et Chaima

Je remercie chaleureusement mes cousines Hanna, Rayen et Chaima pour leur affection, leur présence et leur soutien tout au long de mon parcours. Vos encouragements, vos mots bienveillants et vos sourires m'ont accompagnée dans les moments de fatigue comme dans les instants de joie. Merci d'avoir été là, à votre manière, tout au long de cette aventure.

À ma meilleure amie Nada

À toi, ma précieuse Nada…Merci ne sera jamais un mot assez fort pour exprimer toute ma gratitude envers toi. Tu as été bien plus qu'une amie durant ce parcours : un véritable pilier, une confidente, une lumière dans les moments sombres.

Ta présence constante, ton écoute sans jugement, ton énergie positive et ton amour sincère m'ont portée quand j'avais du mal à avancer. Tu as su sécher mes larmes, apaiser mes doutes, célébrer mes petites victoires et croire en moi même quand je n'y arrivais plus.

Merci pour chaque message d'encouragement, chaque éclat de rire, chaque silence réconfortant. Merci d'avoir été là, sans condition, avec ton cœur immense.

Ton amitié est l'un des plus beaux cadeaux de cette aventure, et je mesure chaque jour la chance que j'ai de t'avoir dans ma vie.

Un merci spécial à ceux qui m'ont aidée en silence

À ceux qui, dans l'ombre, ont déposé un mot, un geste, une prière, une pensée bienveillante sur mon chemin…

À ceux dont l'aide, discrète mais précieuse, a allégé mes fardeaux sans jamais chercher de reconnaissance…

Merci.

Votre présence silencieuse, vos actions cachées, vos attentions invisibles ont eu un impact réel et profond.

Parfois, un simple regard, un sourire, ou une main tendue sans rien dire a suffi à me faire avancer.

Ce remerciement vous est dédié, à vous qui avez contribué à ma réussite avec humilité et cœur.

Je ne vous oublierai jamais.

RÉSUMÉ

L'identification précoce du cancer à partir des données transcriptomiques constitue un enjeu majeur en bioinformatique et en médecine de précision. Dans cette étude, nous avons exploré l'efficacité des réseaux de neurones convolutifs unidimensionnels (Conv1D) pour la classification binaire des profils d'expression génique, en distinguant les échantillons cancéreux des échantillons normaux. Les données utilisées proviennent d'un fichier annoté nommé LUNG_cancer_labeled_5000.csv, contenant des profils transcriptomiques préalablement étiquetés.

Le modèle développé a été entraîné sur 80 % des données et testé sur les 20 % restantes. Il repose sur une architecture composée de couches Conv1D avec activation ReLU, couches de MaxPooling, suivies d'une couche dense et d'un Dropout pour limiter le surapprentissage. Les performances ont été évaluées à l'aide de métriques classiques : précision, rappel, F1-score, matrice de confusion et rapport de classification.

Les résultats obtenus sont très prometteurs, avec une précision de 99,00 % sur l'ensemble de test, une matrice de confusion montrant une excellente capacité de détection des deux classes, et un déséquilibre minimal entre faux positifs et faux négatifs. Ces performances soulignent le potentiel des Conv1D pour des applications biomédicales concrètes.

Des perspectives d'amélioration ont été proposées, notamment l'augmentation des données, l'optimisation des hyperparamètres, l'intégration d'outils d'explicabilité et le recours à des approches d'apprentissage ensembliste. Ces pistes ouvrent la voie vers un déploiement futur dans des environnements cliniques de dépistage assisté par intelligence artificielle.

Mots-clés : Apprentissage profond, Conv1D, Expression génique, Classification binaire, Cancer du poumon, Transcriptomique, Bioinformatique, Réseaux neuronaux, Intelligence artificielle, Dépistage précoce.

ABSTRACT

Early detection of cancer through transcriptomic data analysis is a key challenge in bioinformatics and precision medicine. In this study, we investigate the effectiveness of one-dimensional convolutional neural networks (Conv1D) for binary classification of gene expression profiles, aiming to distinguish cancerous samples from normal ones. The dataset used, LUNG_cancer_labeled_5000.csv, contains pre-labeled gene expression profiles derived from biological samples.

The model was trained on 80% of the data and evaluated on the remaining 20%. The architecture consists of stacked Conv1D layers with ReLU activation functions, followed by MaxPooling layers, a fully connected dense layer, and a Dropout layer to prevent overfitting. Model performance was assessed using common classification metrics, including accuracy, recall, F1-score, confusion matrix, and classification report.

The experimental results are highly promising, with the model achieving 99.00% accuracy on the test set. The confusion matrix shows excellent class separation, with very few false positives and false negatives. These results demonstrate the model's robustness and its potential for real-world application in cancer detection using transcriptomic signatures.

Several directions for improvement are proposed, such as increasing dataset diversity, optimizing hyperparameters using advanced search techniques, applying feature selection and dimensionality reduction methods, and incorporating explainable AI tools. These perspectives contribute to paving the way toward clinical deployment of AI-assisted diagnostic tools based on gene expression analysis.

Keywords: Deep learning, Conv1D, Gene expression, Binary classification, Lung cancer, Transcriptomics, Bioinformatics, Neural networks, Artificial intelligence, Early detection.

الملخص

يمثّل الاكتشاف المبكر للسرطان اعتمادًا على بيانات التعبير الجيني تحديًا رئيسيًا في مجالي المعلوماتية الحيوية والطب الدقيق. في هذه الدراسة، قمنا بدراسة فعالية الشبكات العصبية الالتفافية أحادية البعد (Conv1D) في التصنيف الثنائي لملفات التعبير الجيني، بهدف التمييز بين العينات السرطانية والعينات السليمة. تم استخدام مجموعة بيانات تحمل اسم LUNG_cancer_labeled_5000.csv، وتحتوي على ملفات تعبير جيني معنونة مسبقًا.

تم تدريب النموذج باستخدام 80٪ من البيانات، بينما خُصّصت نسبة 20٪ للاختبار. تتكوّن بنية النموذج من طبقات Conv1Dمتتالية مفعّلة بدالةReLU، متبوعة بطبقاتMaxPooling، ثم طبقة كثيفة (Dense) وطبقة Dropout لتقليل احتمال فرط التعلّم. تم تقييم أداء النموذج باستخدام مقاييس شائعة مثل الدقة، والاسترجاع، ودرجةF1 ، ومصفوفة الالتباس، وتقرير التصنيف.

أظهرت النتائج فعالية واعدة للنموذج، حيث بلغ معدل الدقة 99.00٪ على مجموعة الاختبار، كما أظهرت مصفوفة الالتباس قدرة عالية على التمييز بين الفئتين مع عدد قليل من الإيجابيات والسلبيات الكاذبة. تؤكد هذه النتائج متانة النموذج وفعاليته في التطبيقات الواقعية للكشف عن السرطان باستخدام التوقيعات النسخية.

تم اقتراح عدة آفاق لتطوير العمل، مثل توسيع مجموعة البيانات، وتحسين المعاملات (Hyperparameters)، واستخدام تقنيات الذكاء الاصطناعي القابل للتفسير، ودمج طرق تعلم جماعي (Ensemble Learning). تساهم هذه المقترحات في التمهيد لتطبيق النموذج سريريًا ضمن أدوات تشخيص تعتمد على الذكاء الاصطناعي

الكلمات الرئيسية: التعلّم العميق، الشبكات الالتفافية Conv1D، التعبير الجيني، التصنيف الثنائي، سرطان الرئة، النسخ الجيني، المعلوماتية الحيوية، الشبكات العصبية، الذكاء الاصطناعي، الكشف المبكر

LISTES DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

- ✓ ADN : Acide Désoxyribonucléique
- ✓ AI : Intelligence Artificielle
- ✓ ALK : Anaplastic Lymphoma Kinase
- ✓ APA : Apprentissage Automatique
- ✓ BRAF : B-Raf Proto-Oncogene, Serine/Threonine Kinase
- ✓ CNN : Convolutional Neural Network (Réseau Neuronal Convolutif)
- ✓ CSV : Comma-Separated Values
- ✓ DL : Deep Learning (Apprentissage profond)
- ✓ EGFR : Epidermal Growth Factor Receptor
- ✓ GPU : Graphics Processing Unit (unité de traitement graphique)
- ✓ KRAS : Kirsten Rat Sarcoma Viral Oncogene Homolog
- ✓ MET : MET Proto-Oncogene, Receptor Tyrosine Kinase
- ✓ ML : Machine Learning (Apprentissage Automatique)
- ✓ MYC : MYC Proto-Oncogene, BHLH Transcription Factor
- ✓ PCR : Polymérase Chain Reaction (réaction de polymérisation en chaîne)
- ✓ RET : REarranged during Transfection Proto-Oncogene
- ✓ RNA : Acide Ribonucléique
- ✓ RNA-seq : Séquençage d'ARN
- ✓ ROS1 : ROS Proto-Oncogene 1, Receptor Tyrosine Kinase
- ✓ TP53 : Tumor Protein p53

TABLE DE MATIÈRE

TABLE DES MATIÈRES

INTRODUCTION

INTRODUCTION

La détection précoce et fiable des cancers constitue un enjeu médical et scientifique majeur, en particulier face à l'augmentation constante de l'incidence de ces maladies à travers le monde. Parmi les types de cancer les plus redoutés figure le cancer du poumon, qui reste l'une des principales causes de mortalité liée au cancer, en raison de son diagnostic souvent tardif.

Le cancer du poumon constitue aujourd'hui l'un des problèmes de santé publique les plus préoccupants au niveau mondial. Selon l'Organisation mondiale de la santé (OMS), il s'agit de l'un des cancers les plus fréquents et la première cause de décès par cancer, aussi bien chez les hommes que chez les femmes. Chaque année, des millions de nouveaux cas sont diagnostiqués à travers le monde, avec un taux de mortalité encore très élevé malgré les progrès réalisés en matière de prévention, de diagnostic et de traitement. Cette gravité s'explique en partie par le caractère insidieux de la maladie, qui évolue souvent de manière silencieuse et n'est diagnostiquée qu'à un stade avancé, lorsque les chances de survie sont considérablement réduites (Word Health Organization, 2025).

Le cancer du poumon se divise en deux grandes catégories histologiques : le cancer bronchique non à petites cellules (CBNPC ou NSCLC, Non-Small Cell Lung Cancer) et le cancer bronchique à petites cellules (CBPC ou SCLC, Small Cell Lung Cancer). Le CBNPC représente environ 85 % des cas et se développe lentement, tandis que le CBPC, plus agressif, évolue rapidement et présente un potentiel métastatique important. Ces deux types présentent des différences cliniques et moléculaires, mais partagent un point commun majeur : leur détection précoce reste un défi majeur, notamment en raison de l'absence de symptômes spécifiques dans les premières phases de développement tumoral (Basumallik & Agarwal, 2025).

Dans ce contexte, le dépistage précoce apparaît comme une priorité absolue en oncologie thoracique. Il est largement reconnu que plus un cancer est détecté tôt, plus les chances de survie et de guérison sont élevées. Or, les méthodes de dépistage actuelles — notamment l'imagerie thoracique, la radiographie ou encore la tomodensitométrie (scanner) — bien qu'efficaces dans certaines conditions, présentent des limites : coût élevé, risques d'irradiation, faux positifs ou faux négatifs, et manque de sensibilité pour les petites lésions. De plus, ces méthodes ne permettent pas de caractériser finement la nature biologique de la tumeur. C'est pourquoi de nouvelles approches, plus sensibles, moins invasives et capables

d’exploiter des informations moléculaires, sont activement explorées dans la recherche actuelle (Amicizia et al., 2023).

Grâce aux avancées récentes dans les domaines de la génomique et de l’intelligence artificielle, de nouvelles approches prometteuses émergent pour améliorer les méthodes de dépistage et de classification des tumeurs à un stade précoce.

Parmi ces approches, l’analyse des données d’expression génique représente une voie prometteuse. Cette technique, issue de la génomique fonctionnelle, permet d’étudier l’activité des gènes dans un échantillon biologique donné, en identifiant ceux qui sont surexprimés ou sous-exprimés dans des cellules tumorales par rapport à des cellules saines. Ces différences d’expression peuvent servir de biomarqueurs pour le diagnostic, la classification ou même le suivi thérapeutique du cancer. Les profils transcriptomiques sont généralement obtenus à partir de plateformes technologiques telles que les microarrays (puces à ADN) ou le séquençage haut débit (RNA-Seq), qui génèrent des données riches mais complexes, nécessitant des méthodes analytiques avancées pour en extraire une information pertinente (« Gene Expression Profiling in Cancer », 2025).

Cependant, l’exploitation efficace de ces données reste un défi en soi. Les données d’expression génique sont souvent très volumineuses, bruitées et de très haute dimension (des milliers de gènes mesurés pour un nombre limité d’échantillons). Ce contexte rend les analyses classiques difficiles et ouvre la voie à l’utilisation de méthodes d’analyse plus puissantes, capables de détecter des patterns subtils et non linéaires dans les données. C’est ici qu’intervient l’intelligence artificielle (IA), et plus particulièrement les méthodes d’apprentissage automatique (machine learning) et d’apprentissage profond (deeplearning), qui ont montré des résultats impressionnants dans divers domaines de la biologie, notamment en classification de données génomiques (Hwang et al., 2024).

L’apprentissage profond, avec ses architectures de réseaux de neurones, permet de modéliser des relations complexes dans les données sans nécessiter d’ingénierie manuelle des variables. Parmi les architectures les plus utilisées, les réseaux de neurones convolutifs (CNN) ont d’abord été développés pour l’analyse d’images, mais leur déclinaison unidimensionnelle (Conv1D) s’est révélée particulièrement adaptée à l’analyse de données de type séquentiel, comme les séries temporelles ou les vecteurs de gènes. Ces modèles peuvent automatiquement apprendre des représentations efficaces à partir des profils

transcriptomiques, ce qui en fait un outil puissant pour la détection ou la classification des cancers à partir des données d'expression (Gunavathi et al., 2021).

Dans ce mémoire, nous nous intéressons à la prédiction du cancer du poumon à partir de données d'expression génique annotées, en utilisant un modèle basé sur un réseau de neurones convolutif unidimensionnel (Conv1D). Ce choix repose sur les performances documentées de ce type de réseau pour les tâches de classification binaire dans des contextes biologiques similaires. Afin d'entraîner et évaluer ce modèle, nous avons utilisé un jeu de données prétraité provenant de la base BioStudies, contenant des profils d'expression génique labellisés en deux classes : "cancer" et "normal". Les outils informatiques mobilisés pour ce travail incluent le langage Python et plusieurs bibliothèques spécialisées comme TensorFlow, Keras, Scikit-learn, Pandas, entre autres.

Ce mémoire est structuré en trois chapitres :

- Chapitre 1 : Ce chapitre expose les bases théoriques de l'étude, en décrivant l'épidémiologie, les types histologiques et les mécanismes moléculaires du cancer du poumon. Il présente également les données d'expression génique, les technologies de profilage (microarrays, RNA-Seq), et l'apport de l'intelligence artificielle, en particulier les CNN 1D, dans l'analyse des données transcriptomiques.
- Chapitre 2 : Ce chapitre décrit le jeu de données utilisé, les outils logiciels et matériels (Python, TensorFlow, Keras, etc.), ainsi que les étapes méthodologiques : prétraitement des données, conception du modèle de réseau de neurones convolutif unidimensionnel (Conv1D), entraînement, et évaluation de ses performances.
- Chapitre 3 : Ce chapitre présente les résultats obtenus (précision du modèle, matrice de confusion, rapport de classification), interprète leur signification, et discute les perspectives d'amélioration futures pour renforcer l'efficacité et la robustesse du modèle en vue d'un déploiement en contexte clinique.

CHAPITRE 1 : FONDEMENTS DE LA CLASSIFICATION DU CANCER DU POUMON

1. Cancer du poumon

Le cancer du poumon constitue l'une des formes de cancer les plus meurtrières dans le monde(*Lung Cancer*, 2025). Il résulte de la transformation maligne de cellules pulmonaires normales en cellules cancéreuses, capables de se multiplier de manière incontrôlée et d'envahir les tissus voisins. Cette maladie se caractérise par une évolution souvent silencieuse, ce qui complique son diagnostic précoce et aggrave son pronostic(Mazzone & Lam, 2022). Malgré les avancées thérapeutiques récentes, son taux de mortalité demeure élevé, en raison d'un repérage tardif et de formes cliniques agressives. Comprendre les particularités de ce cancer, tant sur le plan épidémiologique que biologique, est une étape clé pour envisager des stratégies de dépistage plus efficaces et adaptées aux défis posés par cette pathologie (Zarinshenas et al., 2023).

1.1. Épidémiologie et impact en santé publique

Le cancer du poumon constitue un véritable fardeau sanitaire à l'échelle mondiale, tant par son incidence que par sa létalité. D'après les données du Centre international de recherche sur le cancer (CIRC), il représente environ 11,4 % de l'ensemble des cancers diagnostiqués et 18 % des décès liés au cancer, ce qui en fait le cancer le plus mortel toutes localisations confondues (Li et al., 2023). Sa prévalence varie cependant selon les régions du globe, reflétant des inégalités d'exposition aux facteurs de risque, mais aussi des différences en matière de dépistage, d'accès aux soins, et de sensibilisation du public (Chen et al., 2025).

Dans les pays industrialisés, l'incidence tend à diminuer légèrement chez les hommes, notamment grâce aux politiques antitabac, mais elle augmente chez les femmes en raison d'une consommation de tabac plus tardive mais croissante (Islami et al., 2015). En revanche, dans de nombreux pays à revenu faible ou intermédiaire, on observe une hausse continue du nombre de cas, sans que les infrastructures sanitaires soient toujours en mesure de répondre efficacement à cette augmentation (*How to Transform Lung Cancer Outcomes in LMICs*, 2024). Cela se traduit par une prise en charge souvent tardive et une mortalité proportionnellement plus élevée (Nwagbara et al., 2020).

En Algérie, les données épidémiologiques recueillies par le Registre national du cancer indiquent que le cancer du poumon figure parmi les trois cancers les plus fréquents chez l'homme, avec une prédominance nette dans la tranche d'âge supérieure à 50 ans (*Rapport de*

2024 de l'Institut National de Santé Publique, 2024). Le tabagisme reste le facteur de risque principal, mais il n'est pas le seul (tropicale, 2025). Une part importante des cas est également associée à l'exposition professionnelle à des substances cancérogènes, telles que l'amiante, le chrome ou les poussières de silice, notamment dans les secteurs industriels et du bâtiment. La pollution de l'air urbain, en particulier les particules fines, est aussi identifiée comme un facteur environnemental non négligeable, aggravé par la densité du trafic et la combustion domestique dans certaines zones (Z. Melissa, 2024).

Il convient également de souligner l'émergence de facteurs génétiques et biologiques dans la compréhension des susceptibilités individuelles. Des polymorphismes génétiques, des mutations somatiques ou germinales, ainsi que des antécédents familiaux de cancers respiratoires, sont désormais reconnus comme modulateurs du risque, même en l'absence d'exposition à des facteurs exogènes (Gabriel et al., 2022; Sorscher et al., 2023). Cette diversité étiologique complexifie les stratégies de prévention, qui ne peuvent plus se limiter à la seule lutte contre le tabac (Zhou et al., 2025).

En termes de pronostic, le taux de survie à cinq ans reste particulièrement bas. Il est estimé à environ 19 % à l'échelle mondiale, mais peut chuter sous la barre des 10 % dans certaines régions où les diagnostics sont plus tardifs et les options thérapeutiques limitées. À titre de comparaison, ce taux dépasse 80 % pour les cancers du sein détectés précocement (*Lung Cancer Treatment & Survival Rate | City of Hope*, 2025). Cette situation s'explique en grande partie par l'absence de symptômes spécifiques aux premiers stades de la maladie, mais aussi par le retard dans l'adoption de programmes de dépistage systématique, notamment chez les populations à haut risque (« Lung Cancer Screening », 2025).

Enfin, l'impact socio-économique du cancer du poumon ne se limite pas aux aspects médicaux. La maladie touche souvent des individus en âge actif, entraînant une perte de productivité, un poids financier important pour les familles, et une pression accrue sur les systèmes de santé, en particulier dans les pays en développement où les ressources sont déjà limitées (Gonzalez et al., 2023; *S'attaquer à l'impact du cancer sur la santé, l'économie et la société*, 2024).

Face à ce constat, il devient essentiel d'améliorer les stratégies de prévention, de renforcer les moyens de dépistage précoce, et de développer des outils de diagnostic plus sensibles et plus accessibles (Karimzadeh et al., 2024). L'exploitation des données d'expression génique, combinée aux techniques d'intelligence artificielle, s'inscrit dans cette logique d'innovation,

en offrant de nouvelles perspectives pour une détection plus précoce et plus précise de la maladie (Liu & Yao, 2022).

1.2. Types histologiques du cancer du poumon

La classification histologique du cancer du poumon constitue une étape fondamentale dans l'évaluation clinique de la maladie, car elle conditionne en grande partie le pronostic, le choix thérapeutique et l'évolution clinique. Cette classification repose sur l'examen microscopique des cellules tumorales (Figure 1), permettant de distinguer deux grandes entités aux comportements biologiques très différents (Jelic & MD, s. d.; *Types of Lung Cancer | LUNGevity Foundation*, s. d.) : le cancer bronchique non à petites cellules (CBNPC) et le cancer bronchique à petites cellules (CBPC).

Le cancer bronchique non à petites cellules (CBNPC), également désigné sous l'acronyme NSCLC (Non-Small Cell Lung Cancer), représente la forme la plus fréquente, avec environ 85 % des cas diagnostiqués. Il se divise lui-même en plusieurs sous-types histologiques distincts :

- L'adénocarcinome : le plus répandu, en particulier chez les non-fumeurs, il se développe à partir des cellules glandulaires tapissant les voies respiratoires périphériques. Il est souvent détecté sous forme de nodules périphériques à croissance lente.
- Le carcinome épidermoïde : fréquemment observé chez les fumeurs, il se développe dans les bronches principales et peut provoquer une obstruction bronchique, avec des symptômes respiratoires marqués.
- Le carcinome à grandes cellules : forme moins courante, mais plus agressive, souvent diagnostiquée à un stade avancé.

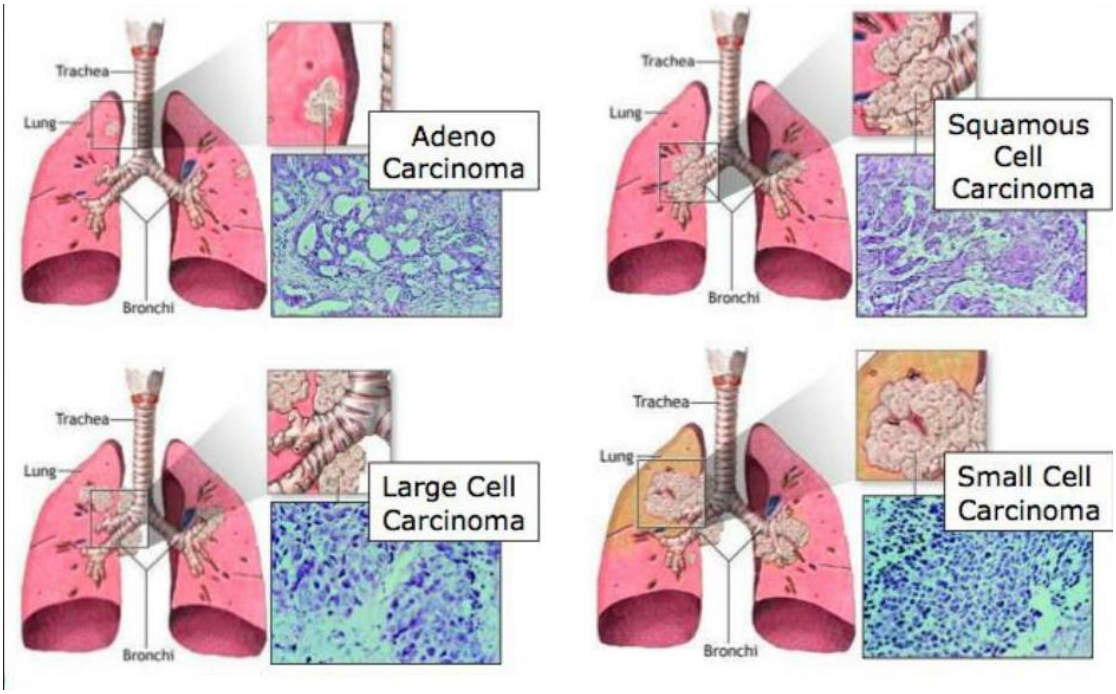

Figure 1 : Représentation schématique et histologique des principaux types de cancer du poumon

Le CBNPC est généralement associé à une progression plus lente et à une meilleure sensibilité aux traitements chirurgicaux et ciblés, notamment lorsqu'il est détecté tôt. Ces dernières années, l'identification de mutations spécifiques (comme EGFR, ALK, KRAS) a permis le développement de thérapies personnalisées, rendant cette forme de cancer plus accessible à une approche de médecine de précision (*Types of Lung Cancer | LUNGevity Foundation*, s. d.).

À l'opposé, le cancer bronchique à petites cellules (CBPC), ou SCLC (Small Cell Lung Cancer), est beaucoup plus rare, mais nettement plus agressif. Il est fortement lié au tabagisme, avec une prévalence nettement plus élevée chez les fumeurs ou anciens fumeurs. Histologiquement, il se caractérise par de petites cellules rondes à noyaux denses, à forte activité mitotique, ce qui traduit une prolifération rapide et un potentiel métastatique élevé. En raison de sa nature systémique dès les stades précoces, le CBPC répond initialement bien à la chimiothérapie et à la radiothérapie, mais présente un taux de rechute particulièrement élevé, réduisant considérablement les chances de survie à long terme (Jelic & MD, s. d.).

Bien que cette distinction en deux grandes catégories soit indispensable dans la pratique clinique, elle ne suffit plus à expliquer l'hétérogénéité biologique des tumeurs pulmonaires. En effet, des patients présentant la même forme histologique peuvent répondre différemment

aux traitements, ou présenter des profils évolutifs divergents. Cela s'explique par des variations moléculaires intra- et inter-tumorales, qui influencent directement la sensibilité aux thérapies, la capacité d'invasion, ou encore le risque de métastases (Cognigni et al., 2025; Y.-C. Fu et al., 2025).

C'est dans ce contexte qu'émerge l'intérêt pour une classification complémentaire basée sur des marqueurs moléculaires, en particulier les profils d'expression génique (Saggi et al., 2024). Ces outils permettent d'aller au-delà de l'apparence microscopique des cellules, en révélant des signatures transcriptomiques spécifiques à certains sous-types ou sous-groupes de tumeurs (Zengin & Önal-Süzek, 2020). En associant l'histologie traditionnelle à des approches de stratification moléculaire, il devient possible de mieux prédire l'évolution de la maladie et de guider plus finement le choix des traitements, notamment dans les cas limite ou peu différenciés (« The Cancer Genome Atlas », 2025).

1.3. Mécanismes moléculaires et profils génétiques

Le développement du cancer du poumon est étroitement lié à l'accumulation progressive de perturbations génétiques et épigénétiques (Figure 2) qui altèrent les mécanismes cellulaires de régulation (ARCAGY-GINECO, 2025a). Ces altérations concernent des fonctions clés de la cellule telles que la prolifération, l'apoptose (mort cellulaire programmée), la différenciation, la réparation de l'ADN ou encore le contrôle du cycle cellulaire. Lorsqu'un déséquilibre se produit dans ces processus, des clones cellulaires anormaux peuvent émerger, résister aux signaux de régulation, et initier un processus tumoral (Wen et al., 2011).

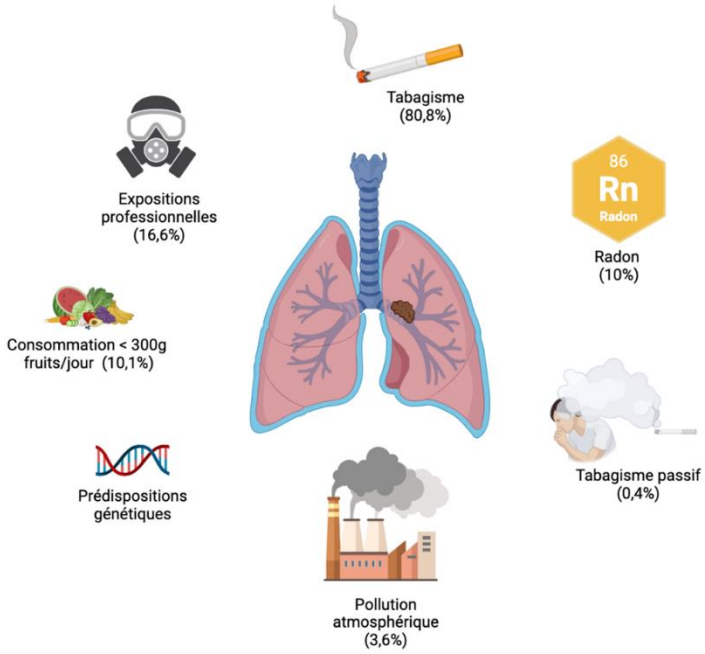

Figure 2 : Principaux facteurs de risque du cancer du poumon

Parmi les altérations les plus fréquentes observées dans le cancer du poumon, on retrouve (Wen et al., 2011) :

- Des mutations ponctuelles dans des gènes codant pour des protéines régulatrices majeures, comme *KRAS* (souvent muté dans les adénocarcinomes), *EGFR* (récepteur de l'EGF), ou *TP53* (gène suppresseur de tumeur impliqué dans la réponse au stress cellulaire) ;
- Des amplifications géniques, telles que celles touchant le gène *MYC*, qui entraînent une surexpression de protéines stimulant la croissance cellulaire ;
- Des délétions, duplications ou translocations chromosomiques, qui réorganisent l'information génétique et peuvent activer des oncogènes ou inactiver des gènes suppresseurs de tumeur ;
- Des modifications épigénétiques, notamment la méthylation aberrante de l'ADN ou des modifications post-traductionnelles des histones, qui influencent l'accessibilité à l'ADN sans altérer la séquence génomique.

Ces anomalies, qu'elles soient génétiques ou épigénétiques, se traduisent par des changements d'expression génique au sein des cellules tumorales, générant ainsi des profils transcriptomiques caractéristiques. Ces signatures transcriptomiques reflètent l'état fonctionnel de la cellule cancéreuse et peuvent être utilisées pour en déduire son agressivité,

son origine, ou encore sa sensibilité à certains traitements. Grâce aux progrès de la biotechnologie, il est aujourd'hui possible d'analyser ces profils à large échelle à l'aide de plateformes à haut débit telles que les microarrays (puces à ADN) ou le séquençage de l'ARN (RNA-Seq). Ces technologies permettent de quantifier simultanément l'expression de milliers de gènes et offrent un aperçu global des processus biologiques perturbés (Jha et al., 2022; Singhal et al., 2008).

L'exploitation de ces données a conduit à la découverte de biomarqueurs moléculaires jouant un rôle central dans le développement de traitements ciblés. Un exemple emblématique est la surexpression du récepteur EGFR, observée dans certains adénocarcinomes pulmonaires, qui a permis le développement d'inhibiteurs spécifiques de tyrosine kinase, capables de bloquer ce signal de croissance anormal. De même, d'autres altérations ciblables comme les réarrangements de *ALK*, *ROS1*, *RET*, ou encore les mutations de *MET* et *BRAF*, ont ouvert la voie à des thérapies personnalisées, souvent mieux tolérées et plus efficaces que les chimiothérapies standards (K. Fu et al., 2022; Wu & Lin, 2022; Yoh, 2019).

Ce changement de paradigme, basé sur la compréhension fine des altérations moléculaires spécifiques à chaque patient, constitue le socle de la médecine de précision. Celle-ci vise à adapter les traitements aux caractéristiques biologiques de chaque tumeur, en s'appuyant sur des données issues de l'expression génique, de la génomique et de la bioinformatique. Dans ce contexte, l'analyse transcriptomique devient un outil diagnostique et pronostique incontournable, capable de guider le clinicien dans le choix du traitement le plus approprié et potentiellement d'améliorer le pronostic du patient (Balan et al., 2023; Li et al., 2024; Pleasance et al., 2022).

1.4. Limites des méthodes classiques de dépistage

Malgré les efforts déployés en santé publique, le dépistage précoce du cancer du poumon demeure un défi majeur, et ce, en grande partie en raison des limites des outils diagnostiques actuellement utilisés. Dans la majorité des cas, le diagnostic intervient à un stade avancé de la maladie, lorsque les symptômes deviennent cliniquement apparents et que les options thérapeutiques sont plus restreintes. Ce retard diagnostique impacte directement le taux de survie, qui pourrait être considérablement amélioré par une détection plus précoce (*Traitement et pronostic du cancer du poumon non à petites cellules*, s. d.; Wilkinson & Lam, 2021).

Les méthodes conventionnelles de dépistage comprennent principalement (*Diagnostic du cancer du poumon – Centres interdisciplinaires d'oncologie*, s. d.; Wilkinson & Lam, 2021) :

- La radiographie thoracique, longtemps utilisée comme outil de première intention. Bien qu'elle soit peu coûteuse et accessible, elle manque de sensibilité, notamment pour les petites lésions périphériques ou les tumeurs précoces. Elle est donc peu fiable pour un dépistage systématique à grande échelle.
- La tomodensitométrie (TDM) ou scanner thoracique, particulièrement la version à faible dose (low-dose CT scan), s'est imposée comme l'outil le plus performant pour le dépistage des sujets à haut risque, notamment les fumeurs âgés de plus de 55 ans. Elle permet de visualiser des nodules non visibles à la radiographie. Cependant, son utilisation soulève des problèmes de coût, d'exposition répétée aux rayonnements, et de spécificité limitée, avec un taux non négligeable de faux positifs, conduisant à des examens complémentaires inutiles, voire invasifs.
- La cytologie des expectorations et les biopsies bronchiques ou pulmonaires sont des méthodes permettant une confirmation histologique du diagnostic. Toutefois, elles sont rarement utilisées en première ligne, car elles nécessitent un acte invasif, sont coûteuses et techniquement exigeantes, et sont généralement réalisées lorsque la suspicion de malignité est déjà forte.

Au-delà de leur caractère technique, ces approches classiques présentent une limite fondamentale : elles sont essentiellement morphologiques et ne permettent pas de détecter les anomalies moléculaires précoces qui précèdent souvent l'apparition de lésions visibles. Elles ne tiennent pas compte de l'hétérogénéité biologique des tumeurs ni de leurs caractéristiques génétiques ou transcriptomiques, ce qui limite leur capacité à stratifier les patients selon leur profil de risque ou leur réponse attendue aux traitements (Imyanitov et al., 2024; Patharia et al., 2024; Wang et al., 2024).

En outre, la précision diagnostique de ces méthodes est loin d'être parfaite. Des faux positifs peuvent conduire à des interventions inutiles sur des lésions bénignes (granulomes, infections, nodules inflammatoires), tandis que des faux négatifs peuvent retarder la prise en charge de véritables cancers, en particulier ceux situés dans des zones difficilement accessibles ou à croissance lente (*Dépistage Du Cancer Du Poumon*, s. d.; *Dépistage Du Cancer Du Poumon*, s. d.; « (PDF) Imagerie Radiologique et TEP Scanner Dans Les Cancers Du Poumon », 2024).

Face à ces limites, il devient impératif de développer des approches complémentaires, plus sensibles, spécifiques et non invasives, capables de détecter le cancer à un stade précoce, voire préclinique. Parmi les pistes les plus prometteuses figure l'analyse de l'expression génique, qui permet d'identifier des signatures moléculaires précoces de transformation cellulaire. Ces signatures, détectables bien avant l'apparition des signes cliniques ou radiologiques, peuvent fournir des indicateurs fiables d'un processus tumoral en cours.

L'intégration de ces données transcriptomiques dans des modèles d'intelligence artificielle, notamment à l'aide d'algorithmes d'apprentissage automatique ou profond, ouvre la voie à une médecine prédictive. Ces outils sont capables de traiter des volumes importants de données, d'identifier des patterns invisibles à l'œil humain, et de fournir des prédictions robustes sur la présence d'un cancer, sa nature, ou encore sa réponse probable à certains traitements. Ainsi, cette approche combinée représente une avancée stratégique vers un dépistage plus intelligent, mieux adapté à la complexité du cancer du poumon.

2. Les données d'expression génique

L'étude de l'expression des gènes, appelée transcriptomique, occupe aujourd'hui une place centrale dans les recherches biomédicales, notamment dans le domaine de l'oncologie. Grâce aux technologies modernes de biologie moléculaire et aux outils bioinformatiques, il est désormais possible de mesurer et d'analyser l'activité de milliers de gènes en parallèle, à partir d'un simple échantillon biologique. Ces données dites « d'expression génique » reflètent l'état fonctionnel d'une cellule à un instant donné et constituent un outil précieux pour la compréhension des mécanismes pathologiques, le diagnostic, le pronostic, et même la prédiction de la réponse aux traitements (Sun & Chen, 2023; Vekris & Robert, 2005).

2.1. Qu'est-ce que l'expression génique et les données transcriptomiques ?

L'expression génique correspond au processus par lequel l'information contenue dans un gène est utilisée pour produire une molécule fonctionnelle, généralement une protéine (Figure 3). Ce processus comporte plusieurs étapes, dont la transcription, qui est la synthèse d'un brin d'ARN messager (ARNm) à partir de l'ADN. L'ARNm est ensuite traduit en protéine.

L'ensemble des transcrits présents dans une cellule à un moment donné constitue ce que l'on appelle le transcriptome.

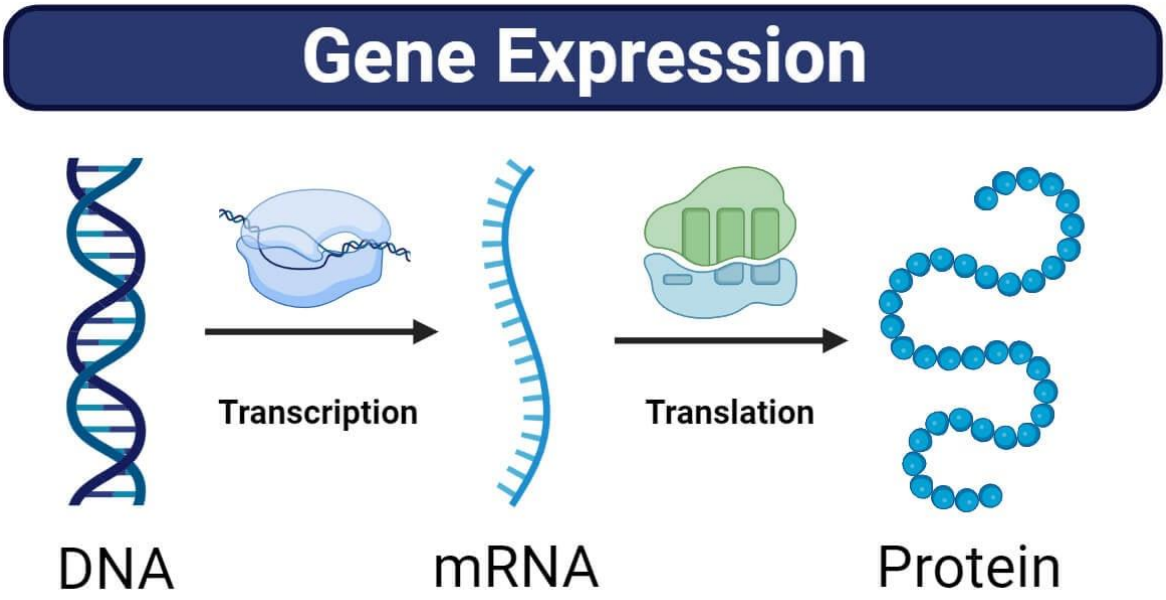

Figure 3 : étapes de l'expression génique : de l'ADN à la protéine

Les données transcriptomiques désignent donc les données quantitatives obtenues à partir de l'analyse de l'ARNm extrait des cellules. Ces données permettent de déterminer quels gènes sont actifs (exprimés) ou inactifs, et dans quelle mesure. Elles sont souvent présentées sous la forme de tableaux contenant l'expression relative ou absolue de chaque gène dans différents échantillons (Alberts et al., 2002).

L'analyse de l'expression génique est particulièrement utile dans le contexte du cancer, car les cellules tumorales présentent des profils d'expression très différents de ceux des cellules saines. Certains gènes peuvent être surexprimés (gènes promoteurs de tumeurs) ou au contraire réprimés (gènes suppresseurs de tumeurs), offrant ainsi une signature moléculaire unique à chaque type tumoral (Alberts et al., 2002).

Cette signature peut ensuite être exploitée pour (Šutić et al., 2024) :

- Identifier des biomarqueurs spécifiques ;
- Développer des tests diagnostiques ou pronostiques ;
- Classer les cancers selon leur origine ou leur stade ;
- Orienter le choix des traitements en fonction du profil transcriptomique du patient.

2.2. Technologie de mesure utilisée : les Microarrays

L'obtention de données d'expression génique repose sur des technologies de haut débit capables de mesurer simultanément l'abondance de milliers de transcrits dans un échantillon biologique. Dans le cadre du présent travail, les données analysées ont été obtenues à l'aide de la technologie des microarrays (Figure 4), ou puces à ADN, qui reste l'une des méthodes les plus répandues et les mieux établies pour l'étude du transcriptome, en particulier dans les grandes études cliniques et les bases de données publiques (Alberts et al., 2002).

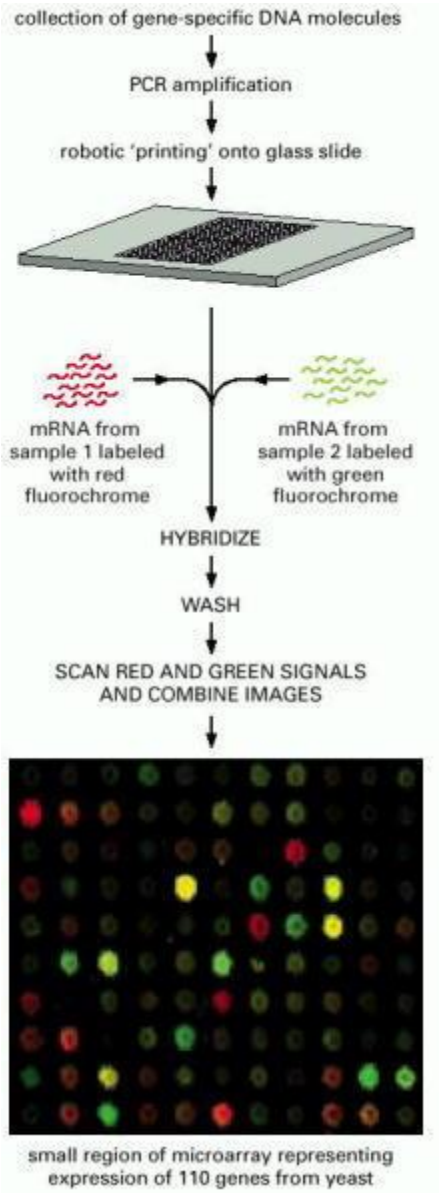

Figure 4 : Schéma du fonctionnement d'une puce à ADN (microarray)

Les microarrays consistent en des supports solides (souvent des lames de verre ou des membranes en silicone) sur lesquels sont fixées de manière ordonnée des sondes d'ADN complémentaires correspondant à des séquences de gènes d'intérêt. Lors de l'expérimentation, l'ARN extrait d'un tissu (sain ou tumoral) est converti en ADN complémentaire (ADNc), puis marqué par un colorant fluorescent. Ce mélange est ensuite hybridé sur la puce. Les sondes fixées sur la puce capturent les brins complémentaires présents dans l'échantillon, et l'intensité du signal fluorescent émis est proportionnelle à la quantité d'ARNm d'origine, permettant ainsi une estimation quantitative de l'expression de chaque gène (*Microarray | Learn Science at Scitable*, s. d.; Tarca et al., 2006).

Parmi les plateformes les plus largement utilisées, Affymetrix est particulièrement réputée pour la précision de ses sondes et la couverture étendue du génome humain. Les puces Affymetrix HG-U133 Plus 2.0, notamment, sont fréquemment employées dans les études oncologiques, car elles permettent de mesurer l'expression de plus de 47 000 transcrits. D'autres fournisseurs comme Agilent Technologies ou Illumina proposent également des solutions robustes et reproductibles pour les analyses d'expression génique (*GeneChip™ Human Genome U133 Plus 2.0 Array*, s. d.).

Les avantages majeurs des microarrays incluent :

- Une technologie éprouvée, utilisée depuis les années 1990, avec un vaste historique de données publiées (« Transcriptomique », 2025).
- Un coût relativement accessible par échantillon, surtout en comparaison avec les technologies de séquençage plus récentes (Malone & Oliver, 2011).
- Une standardisation élevée, permettant une reproductibilité inter-laboratoire appréciable et facilitant les études multicentriques ou méta-analyses (Jaluria et al., 2007).

Cependant, cette technologie présente également certaines limites techniques qui doivent être prises en compte lors de l'interprétation des résultats (Bumgarner, 2013; Chauhan, 2019) :

- Les microarrays ne permettent la détection que des gènes connus et préalablement sélectionnés, ce qui exclut la découverte de nouveaux transcrits ;
- Ils sont parfois affectés par des bruits de fond et des effets d'hybridation croisée, qui peuvent nuire à la précision des mesures pour des séquences très similaires (comme les pseudogènes ou familles multigéniques) ;

- Leur plage dynamique est plus restreinte que celle des méthodes de séquençage, limitant la capacité à détecter des différences d'expression très subtiles ou des transcrits faiblement exprimés.

Malgré ces limites, les microarrays conservent toute leur pertinence dans de nombreux projets, notamment ceux qui s'appuient sur des données publiques déjà disponibles, bien annotées et facilement exploitables dans des contextes bioinformatiques. Ils demeurent un outil fiable pour identifier des signatures d'expression associées à des états pathologiques, y compris dans le cadre du cancer du poumon, et constituent une base solide pour le développement de modèles de prédiction ou de classification basés sur l'intelligence artificielle (Chauhan, 2019).

2.3. Les bases de données transcriptomiques disponibles

Pour faciliter la recherche et la reproductibilité des analyses, de nombreuses bases de données publiques mettent à disposition des jeux de données transcriptomiques issus de diverses pathologies, dont le cancer du poumon. Ces ressources sont devenues indispensables pour les chercheurs qui souhaitent entraîner ou tester des modèles prédictifs à partir de données réelles.

2.3.1. GEO (Gene Expression Omnibus)

La base GEO est l'une des plus anciennes et des plus utilisées. Hébergée par le National Center for Biotechnology Information (NCBI), elle contient des centaines de milliers d'expériences transcriptomiques provenant de microarrays et de RNA-Seq (*Frequently Asked Questions - GEO - NCBI*, s. d.).

Chaque jeu de données est annoté avec des métadonnées cliniques, biologiques et techniques, ce qui facilite les recherches ciblées. Il est également possible de télécharger les matrices d'expression normalisées, prêtes à être analysées statistiquement ou intégrées dans des modèles d'IA (Clough et al., 2024).

2.3.2. BioStudies

BioStudies, base qui héberge le dataset utilisé dans le présent mémoire (E-MTAB-3732), est une plateforme récente qui regroupe les différentes modalités de données expérimentales – transcriptomiques, protéomiques, métabolomiques – et permet d'y associer des documents complémentaires, des pipelines, ou des scripts d'analyse.

Son avantage réside dans la structuration et la traçabilité des données. Les chercheurs peuvent ainsi reproduire intégralement les expériences ou les enrichir en y associant des approches nouvelles, comme l'apprentissage automatique (BioStudies, s. d.).

2.3.3. The Cancer Genome Atlas (TCGA)

Bien que non utilisé directement ici, il est utile de mentionner TCGA, qui constitue une référence mondiale en matière de données génomiques et transcriptomiques sur le cancer. Cette initiative du NIH a permis de générer des profils d'expression pour des milliers d'échantillons tumoraux, accompagnés de données cliniques très détaillées. TCGA est largement exploité dans les études de prédiction et de modélisation du cancer à l'aide de l'intelligence artificielle (*Using TCGA - NCI*, 2019).

2.4. Études ayant utilisé les données transcriptomiques pour la détection du cancer

De nombreuses publications ont démontré le potentiel des données transcriptomiques pour améliorer le diagnostic du cancer, notamment dans des situations où les méthodes conventionnelles sont insuffisantes. Voici quelques exemples marquants.

2.4.1. Détection précoce du cancer du poumon par signatures transcriptomiques

Une étude de (Huang et al., 2018) a exploré l'utilisation de l'apprentissage automatique, notamment des machines à vecteurs de support (SVM), pour analyser les données de séquençage de l'ARN (RNA-Seq) de patients atteints de cancer du poumon non à petites cellules (NSCLC). Les auteurs ont identifié des gènes dont l'expression était significativement altérée dans les échantillons tumoraux et ont construit un modèle de classification basé sur SVM, atteignant une précision notable dans la détection de la maladie.

2.4.2. Application du deeplearning sur les données RNA-Seq

Dans une étude publiée par (Mohammed et al., 2021), une architecture de réseau de neurones convolutifs unidimensionnels (1D-CNN) a été utilisée pour classer des profils RNA-Seq issus

du TCGA. Le modèle a permis de distinguer efficacement plusieurs types de cancer, y compris le cancer du poumon, en identifiant automatiquement les gènes les plus discriminants. Les résultats ont montré que le deeplearning peut surpasser les méthodes classiques de classification, tout en s'adaptant à des jeux de données bruyants ou déséquilibrés.

2.4.3. Études comparatives sur les technologies (microarray vs RNA-Seq)

Des travaux comme celui de (Kim et al., 2024) ont comparé les performances de modèles prédictifs construits à partir de données issues de microarrays et de RNA-Seq. Bien que les deux approches donnent des résultats comparables pour certains types de cancer, le RNA-Seq offre une meilleure sensibilité dans la détection de transcrits rares et une résolution plus fine des isoformes, ce qui améliore la précision des modèles.

2.4.4. Intégration de données d'expression dans la médecine de précision

Enfin, plusieurs équipes ont intégré les données d'expression dans des algorithmes de stratification thérapeutique, permettant d'assigner un traitement personnalisé en fonction du profil transcriptomique. Ces travaux, notamment ceux du consortium Lung-MAP(*Lung-MAP Clinical Trial - NCI*, 2014), montrent que les signatures d'expression peuvent être utilisées non seulement pour détecter la maladie, mais aussi pour prédire la réponse à des traitements ciblés ou immunothérapies.

3. Utilisation de l'intelligence artificielle en cancérologie

Les progrès récents de l'informatique et de la biologie ont ouvert la voie à l'intégration de l'intelligence artificielle (IA) dans la recherche biomédicale, en particulier dans le domaine du cancer. Grâce à sa capacité à analyser rapidement de grandes quantités de données complexes, l'IA constitue un outil prometteur pour améliorer le diagnostic, affiner les prédictions pronostiques, et personnaliser les traitements (*Traitement personnalisé du cancer avec l'IA : s'orienter dans le paysage génomique*, s. d.). Dans cette section, nous introduisons brièvement les concepts clés de l'intelligence artificielle appliquée aux données biologiques,

en soulignant ses avantages, ainsi que son application spécifique à l'analyse des profils d'expression génique.

3.1. Définitions de l'intelligence artificielle, du machine learning et du deeplearning

L'intelligence artificielle (IA) désigne un ensemble de techniques informatiques visant à reproduire certaines fonctions cognitives humaines, comme la reconnaissance de motifs, l'apprentissage à partir de l'expérience, la prise de décision ou la prédiction. Contrairement aux algorithmes classiques qui suivent des instructions fixes, l'IA est capable de s'adapter aux données, d'apprendre des exemples, et de générer des réponses optimisées, même face à des situations nouvelles (*Qu'est-ce que l'intelligence artificielle (IA) et pourquoi est-elle importante | NetApp*, s. d.).

Au sein de l'IA, on distingue une sous-discipline appelée apprentissage automatique ou machine learning (ML). Le ML consiste à entraîner des algorithmes sur des ensembles de données existants afin qu'ils apprennent à reconnaître des schémas ou à faire des prédictions sans être explicitement programmés pour chaque tâche. Ces modèles sont ensuite capables de faire des généralisations à partir de nouvelles données. Les algorithmes les plus connus en ML incluent les arbres de décision, les forêts aléatoires, les machines à vecteurs de support (SVM) ou encore les réseaux de neurones artificiels (Robert, 2020b).

Un sous-ensemble du machine learning, appelé deep learning (apprentissage profond) (Figure 5), repose sur l'utilisation de réseaux de neurones profonds (deep neural networks), composés de plusieurs couches de traitement interconnectées. Ces architectures sont capables d'extraire automatiquement des caractéristiques pertinentes à différents niveaux de complexité, sans nécessiter une phase de sélection manuelle des variables. Le deep learning est particulièrement performant dans les domaines à forte dimensionnalité, comme l'imagerie médicale ou l'analyse des données transcriptomiques, où les relations entre les variables sont complexes, non linéaires, et souvent inconnues a priori (Robert, 2020a).

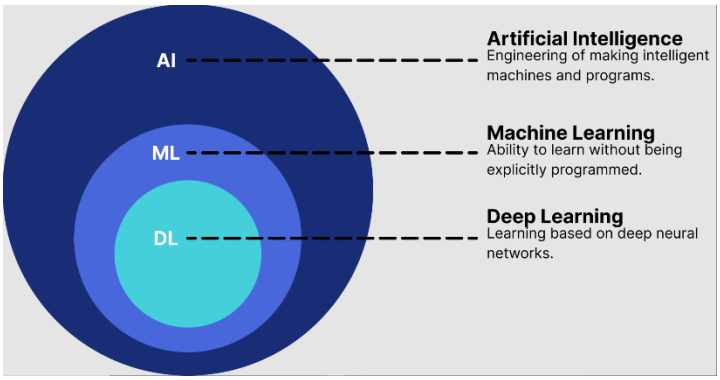

Figure 5: De l'Intelligence artificielle au Deep Learning

3.2. Pourquoi utiliser l'IA en cancérologie ?

L'application de l'IA en cancérologie se justifie par les défis spécifiques liés à cette discipline, notamment (ARCAGY-GINECO, 2025b; *L'IA transforme le combat contre le cancer avec des traitements personnalisés*, 2024) :

- La complexité biologique des cancers, qui impliquent des altérations multiples et souvent interconnectées.
- La variabilité interindividuelle importante, même au sein d'un même type histologique.
- Le volume croissant de données biomédicales disponibles, qu'il s'agisse de données cliniques, génétiques, transcriptomiques ou d'imagerie.

Dans ce contexte, les approches traditionnelles d'analyse de données atteignent rapidement leurs limites. L'IA, en revanche, permet de (ARCAGY-GINECO, 2025b; *L'IA transforme le combat contre le cancer avec des traitements personnalisés*, 2024) :

- Automatiser le traitement de jeux de données volumineux, en réduisant considérablement le temps nécessaire à l'analyse ;
- Identifier des relations complexes ou des motifs cachés que l'œil humain ou les approches statistiques classiques ne pourraient pas détecter ;
- Améliorer la précision du diagnostic ou de la classification, en se basant sur des signatures moléculaires plus fines que les critères morphologiques classiques ;

- Aider à la stratification des patients, en regroupant ceux qui partagent des profils biologiques ou moléculaires similaires, ce qui est essentiel pour la médecine personnalisée ;
- Prédire la réponse aux traitements, grâce à des modèles capables de repérer des corrélations entre les profils d'expression génique et l'efficacité de certaines thérapies.

En d'autres termes, l'IA agit comme un catalyseur dans l'extraction de connaissances à partir des données biologiques complexes, et permet de transformer les données brutes en décisions cliniques éclairées.

4. Les modèles CNN appliqués à la bioinformatique

L'intelligence artificielle, et plus spécifiquement l'apprentissage profond, a profondément transformé l'analyse des données biologiques. Parmi les nombreuses architectures de deeplearning disponibles, les réseaux de neurones convolutifs (Convolutional Neural Networks – CNN) se sont d'abord illustrés dans le domaine de la vision par ordinateur, avant de trouver de multiples applications en bioinformatique, notamment dans le traitement des données transcriptomiques (*La Bioinformatique Connaît Une Innovation Significative Grâce à l'IA et à l'apprentissage Automatique | HackerNoon*, s. d.). Cette section propose une introduction simplifiée aux CNN, avant d'explorer leur adaptation en version 1D pour les profils d'expression génique, ainsi que les études qui en démontrent l'efficacité dans la classification binaire entre tissus cancéreux et non cancéreux.

4.1. Introduction aux réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) sont une architecture de deeplearning initialement conçue pour le traitement des images 2D, mais dont les principes peuvent être généralisés à d'autres types de données. Leur particularité réside dans leur capacité à extraire automatiquement des caractéristiques pertinentes à partir de données brutes, en appliquant des filtres de convolution qui détectent des motifs locaux dans les entrées (*Introduction to Convolution Neural Network*, 2025).

Un CNN est généralement composé de plusieurs types de couches (Figure 6) (*Introduction to Convolution Neural Network*, 2025) :

- Couches convolutives, où les filtres glissent sur les données pour produire des cartes d'activation mettant en évidence les motifs détectés ;
- Couches de regroupement (pooling), qui réduisent la dimension des données et rendent le modèle plus robuste aux petites variations ;
- Couches entièrement connectées, qui interprètent les motifs extraits pour produire une prédiction finale (exemple : classification binaire).

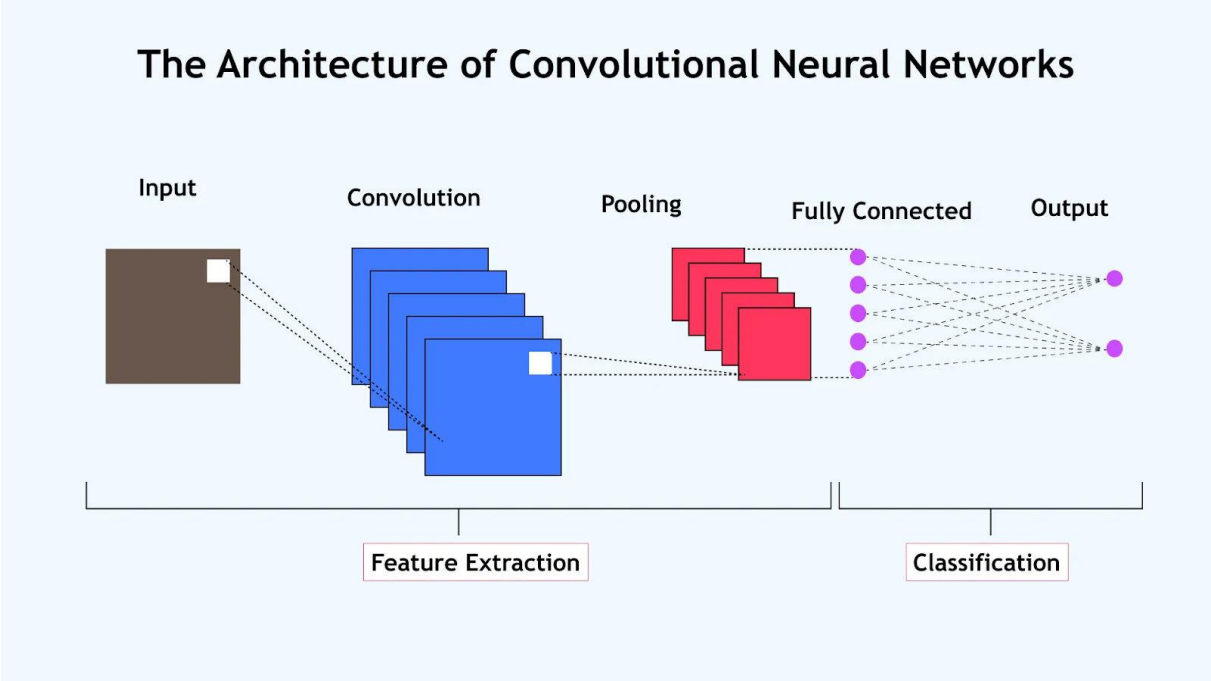

Figure 6 : Structure d'un CNN

Le succès des CNN dans des domaines comme la reconnaissance faciale, la détection d'objets ou l'analyse d'images médicales repose sur leur capacité à (*Introduction aux réseaux neuronaux convolutifs (CNN)*, 2025) :

- Apprendre des représentations hiérarchiques (du motif local simple à la structure globale complexe) ;
- Réduire le besoin de prétraitement manuel ou de sélection de variables ;
- S'adapter à différentes échelles et structures de données.

Dans le domaine biologique, ces propriétés peuvent être exploitées pour identifier des motifs caractéristiques d'un état pathologique, tels que les altérations dans les profils d'expression génique entre échantillons sains et tumoraux.

4.2. Pourquoi utiliser des CNN 1D pour les données d'expression génique ?

Contrairement aux images qui sont représentées sous forme de matrices 2D (pixels en hauteur et largeur), les profils d'expression génique sont des vecteurs unidimensionnels. Chaque gène est représenté par une valeur continue (intensité d'expression), et tous les gènes sont ordonnés dans un tableau par leur position ou leur fonction.

C'est pourquoi, dans ce contexte, on utilise des réseaux de neurones convolutifs unidimensionnels (Conv1D), spécialement conçus pour traiter ce type de données. Dans un modèle Conv1D (Afify et al., 2024; Mohamed et al., 2023) :

- Les filtres convolutifs glissent le long du vecteur de gènes pour capturer des motifs locaux d'expression ;
- Ces motifs peuvent représenter des interactions ou co-régulations entre gènes contigus ou appartenant à une même voie biologique ;
- Le modèle apprend à associer ces motifs à une classe cible, par exemple "cancer" ou "normal".

Les Conv1D sont donc particulièrement adaptés pour détecter des motifs régulatoires dans des séries biologiques où l'ordre ou la proximité des variables (gènes) peut porter une signification fonctionnelle.

Plusieurs raisons justifient leur popularité dans l'analyse des données transcriptomiques (Parisapogu et al., 2021):

- Ils sont moins coûteux en calcul que les CNN 2D, car les opérations se font sur une seule dimension.
- Ils nécessitent moins de paramètres que les réseaux entièrement connectés classiques, ce qui réduit le risque de surapprentissage.
- Ils peuvent être entraînés sur des jeux de données de taille modérée, tout en conservant une bonne capacité de généralisation si les filtres sont bien conçus.

Dans le cadre de la classification binaire (cancer vs normal), les CNN 1D permettent de repérer des signatures d'expression génique caractéristiques des tissus tumoraux, sans nécessiter une connaissance préalable des gènes d'intérêt.

4.3. Avantages et limites des CNN pour les données biologiques

Comme tout modèle, les CNN 1D présentent à la fois des avantages significatifs et des limitations à prendre en compte, en particulier lorsqu'ils sont appliqués à des données transcriptomiques.

Avantages (Dey et al., 2022; Kakati et al., 2022) **:**

- Détection automatique des motifs pertinents : les filtres de convolution capturent des relations locales entre gènes sans intervention manuelle.
- Robustesse aux redondances : les CNN tolèrent les corrélations entre variables, fréquentes dans les profils d'expression.
- Réduction du nombre de paramètres : en comparaison avec un MLP classique, un CNN nécessite moins de poids à apprendre, ce qui limite les risques de surapprentissage.
- Exploitation des structures locales : certains gènes co-régulés peuvent être détectés comme blocs par le réseau.
- Bonne performance avec des données de taille moyenne, typiques des études bioinformatiques utilisant des données de microarrays.

Limites (Afify et al., 2024; Kaissar et al., 2025) **:**

- Besoin d'un prétraitement standardisé : l'ordre des gènes dans le vecteur d'entrée influence les motifs détectés, ce qui nécessite une normalisation rigoureuse des données (échelle, tri, imputation).
- Risque de surapprentissage : en particulier lorsque le nombre d'échantillons est faible par rapport au nombre de gènes. Des techniques comme le dropout ou la régularisation L2 sont alors indispensables.

- Difficulté d'interprétation biologique directe : bien que les CNN apprennent des motifs utiles, leur signification biologique n'est pas toujours claire sans outils d'explicabilité.
- Sensibilité au déséquilibre de classes : comme beaucoup de modèles supervisés, les CNN 1D peuvent être biaisés si les données présentent un déséquilibre important entre classes (par exemple, beaucoup plus de données "cancer" que "normal").
- Optimisation délicate des hyperparamètres : le choix du nombre de filtres, de la taille du kernel, du taux de dropout, ou de la profondeur du réseau peut grandement influencer les performances.

Les réseaux de neurones convolutifs unidimensionnels (CNN 1D) offrent un cadre particulièrement adapté à l'analyse des données transcriptomiques, notamment pour des tâches de classification binaire comme la détection du cancer. Leur capacité à extraire automatiquement des motifs d'expression caractéristiques, tout en réduisant la complexité du modèle, en fait un outil puissant dans les mains des bioinformaticiens et data scientists. Toutefois, comme tout outil, leur efficacité dépend fortement de la qualité des données d'entrée, du prétraitement, et de la gestion des biais éventuels.

L'intégration des CNN 1D dans des pipelines d'analyse standardisés, associée à des outils d'interprétabilité, ouvre des perspectives intéressantes pour améliorer le diagnostic, affiner la stratification des patients, et renforcer la place de la transcriptomique dans la médecine personnalisée.

CHAPITRE 2 : MATÉRIEL ET MÉTHODES

1. Matériel

1.1. Dataset

Dans le cadre de cette recherche, les données exploitées proviennent de la plateforme BioStudies, reconnue pour la qualité et la richesse de ses ressources biologiques. Le jeu de données utilisé est constitué de plusieurs fichiers complémentaires, chacun jouant un rôle déterminant dans la conduite de l'étude (Tableau 1).

Description des fichiers :

Tableau 1 : Présentation des fichiers composant le Dataset

Fichiers :	Taille :	Type :	Description :
processedMatrix.Aurora.july2015.txt	25.75 Go	Données prétraitées	Ce fichier représente la base principale de l'analyse. Il regroupe l'ensemble des informations transformées issues de l'étude, incluant les mesures relatives aux échantillons ainsi qu'aux variables biologiques analysées.
E-MTAB-3732.idf.txt	2 Ko	Fichier IDF (Investigation Description Format)	Ce fichier contient les détails sur la structure expérimentale de l'étude ainsi que les métadonnées y afférentes. Il permet de contextualiser les données et de comprendre le cadre expérimental global.
E-MTAB-3732.sdrf.txt	10.1 Mo	Fichier SDRF (Sample and Data Relationship Format)	Ce fichier joue un rôle clé dans l'annotation des échantillons. Il fournit les étiquettes et les liens nécessaires pour relier les données à leurs conditions expérimentales, facilitant ainsi leur catégorisation et analyse.

1.2. Configuration matérielle et logicielle

Les expérimentations ont été effectuées sur une machine disposant des ressources matérielles nécessaires pour exécuter des algorithmes d'apprentissage profond de manière efficace (Tableau 2). L'ordinateur utilisé intègre un processeur AMD Ryzen 5 3600x, accompagné de 48 Go de mémoire vive (RAM), et d'une carte graphique NVIDIA GeForce RTX 3060 Ti avec 8 Go de mémoire vidéo dédiée (VRAM). L'environnement logiciel est basé sur le système d'exploitation Windows 10, en version 64 bits.

Cette configuration a été utilisée pour sa capacité à gérer des charges de travail computationnelles élevées, en particulier lors de la phase d'entraînement du modèle de deeplearning.

Tableau 2 : Les caractéristiques de l'ordinateur utilisé lors de l'apprentissage profond

Composant	**Caractéristiques**
Processeur	AMD Ryzen 5 3600x
RAM	48 Go
GPU	NVIDIA GeForce RTX 3060 Ti – 8 Go de VRAM
Système d'exploitation	Windows 10
Type de système	Système d'exploitation 64 bits

1.3. Logiciels et Bibliothèques

Le développement du projet s'est appuyé sur Python 3.8, un langage de programmation polyvalent et particulièrement prisé dans les domaines de la bioinformatique, de la science des données et de l'intelligence artificielle. Sa syntaxe claire, son extensibilité via des bibliothèques spécialisées, et sa compatibilité multi-plateforme en font un outil de choix pour des applications scientifiques avancées.

Pour garantir une gestion efficace des dépendances et des environnements, l'ensemble du travail a été réalisé à l'aide de la distribution Anaconda, tandis que le codage et l'expérimentation ont été menés dans Jupyter Notebook, une interface interactive permettant d'imbriquer code, commentaires, équations et graphiques dans un même document.

- **Python :**

Créé par Guido van Rossum et publié en 1991, Python s'est imposé comme un langage incontournable dans l'écosystème scientifique et technique. Son approche orientée objet, sa richesse en bibliothèques spécialisées, et son caractère open source en font une solution adaptée aussi bien aux débutants qu'aux chercheurs expérimentés. Il est particulièrement utilisé pour le traitement de données biologiques, l'analyse statistique, la modélisation, et le développement de modèles d'apprentissage automatique.

- **Anaconda :**

Anaconda est une distribution libre intégrant plus de 1 500 packages pour la science des données. Elle offre un environnement complet et stable, incluant des outils comme **Conda** (gestionnaire de paquets et d'environnements), et des bibliothèques populaires telles que **NumPy**, **Pandas** ou **Scikit-learn**. Elle permet d'éviter les conflits entre dépendances, ce qui est essentiel pour la reproductibilité des expériences.

- **Jupyter notebook :**

Jupyter Notebook est une interface web interactive conçue pour la création de documents combinant du code, des visualisations et du texte. Elle est très utilisée en recherche scientifique car elle favorise l'expérimentation rapide, la transparence des analyses, et la communication des résultats. Chaque cellule peut être exécutée indépendamment, facilitant les tests successifs et la traçabilité des traitements.

- **Bibliothèques et outils utilisés**

Divers outils logiciels et bibliothèques ont été mobilisés pour répondre aux exigences du projet, de la manipulation des données à l'entraînement des modèles, en passant par leur évaluation et visualisation :

- **Pandas (v1.2.4)** : pour le traitement et l'analyse de structures de données tabulaires (DataFrame).
- **NumPy (v1.19.5)** : pour les calculs numériques performants sur des tableaux multidimensionnels.
- **Scikit-learn (v0.24.1)** : pour le prétraitement des données, la sélection des variables et l'évaluation des performances.
- **TensorFlow (v2.4.1)** et **Keras (v2.4.3)** : pour la conception, l'entraînement et l'optimisation de modèles d'apprentissage profond.
- **Matplotlib (v3.3.4)** et **Seaborn (v0.11.1)** : pour la création de graphiques clairs et esthétiques.

2. Méthodes

2.1. Prétraitement des données

Le jeu de données utilisé pour cette étude a été importé à partir d'un fichier au format CSV, nommé « LUNG_cancer_labeled_5000.csv ». Celui-ci contient des échantillons biologiques annotés selon deux classes distinctes : "cancer" et "normal", ce qui facilite l'exploitation directe dans un cadre de classification binaire.

Les données ont été chargées dans un DataFrame à l'aide de la bibliothèque pandas, afin de permettre une manipulation efficace. Étant déjà étiqueté selon les deux catégories d'intérêt, aucune opération de filtrage des classes n'a été nécessaire.

Avant d'entamer les phases de modélisation, une vérification de la distribution des classes a été effectuée pour s'assurer d'un équilibre relatif entre les deux groupes. Cette étape est essentielle afin de prévenir les biais dans l'apprentissage supervisé. D'éventuelles valeurs manquantes, doublons ou incohérences ont également été inspectées et traitées si nécessaire, garantissant ainsi la qualité et la fiabilité des données exploitées.

1) Lecture et Conversion de Données CSV :

L'importation du jeu de données s'effectue à l'aide de la bibliothèque pandas, en utilisant la fonction dédiée à la lecture de fichiers CSV. Le fichier LUNG_cancer_labeled_5000.csv est ainsi chargé dans un objet DataFrame, ce qui permet une manipulation aisée des données pour les étapes ultérieures d'analyse et de modélisation (Figure 7).

```
1  import pandas as pd
2
3  df = pd.read_csv("LUNG_cancer_labeled_5000.csv")
4
```

Figure 7 :lecture et chargement d'un fichier CSV dans un DataFrame

2) Analyse de la distribution des classes :

Une étape préliminaire importante dans tout processus de classification consiste à examiner la répartition des classes dans le jeu de données. Dans ce cas, la distribution

des étiquettes présentes dans la colonne label du DataFramedf a été analysée afin d'évaluer l'équilibre entre les classes "cancer" et "normal". Le code utilisé pour effectuer cette opération est présenté dans la Figure 8.

```
category_counts = df['label'].value_counts()
category_counts
```

Figure 8 : Calcul et affichage de la répartition des classes

La méthode value_counts () permet de comptabiliser le nombre d'occurrences de chaque catégorie dans la colonne spécifiée. Le résultat retourné est un objet Series de pandas, où chaque ligne correspond à une catégorie unique accompagnée de son effectif.

L'exécution de ce code produit une sortie textuelle indiquant le nombre total d'échantillons pour chacune des classes présentes dans la colonne label. Une illustration de cette sortie est fournie dans la Figure 9.

```
cancer    773
normal    233
Name: label, dtype: int64
```

Figure 9 : Répartition des Catégories

2.2. Architecture du modèle de réseau de neurones convolutif 1D

Pour cette étude, un réseau de neurones convolutif unidimensionnel (Conv1D) a été implémenté, en raison de son efficacité à détecter des motifs locaux pertinents dans les séries de données, notamment les profils d'expression génique. Ce type d'architecture est particulièrement adapté au traitement de données génomiques où les relations entre gènes peuvent se manifester sous forme de motifs répartis dans les vecteurs d'entrée.

L'entrée du modèle est d'abord redimensionnée via une couche Reshape afin d'adapter la forme des données à (5000, 1), ce qui est requis pour l'application des convolutions unidimensionnelles.

Le cœur du réseau est constitué de trois couches Conv1D, la première utilisant 32 filtres avec un noyau de taille 9, la seconde 64 filtres avec un noyau de taille 7 et la troisième 128 filtres avec un noyau de taille 5. Ces couches utilisent la fonction d'activation ReLU pour introduire de la non-linéarité et améliorer la capacité d'apprentissage du modèle. Chacune de ces couches convolutionnelles est suivie d'une couche de MaxPooling1D, qui permet de réduire la dimensionnalité des cartes de caractéristiques et d'atténuer la complexité computationnelle.

Les sorties des couches convolutives sont ensuite aplaties à l'aide de la couche Flatten, avant d'être transmises à une couche dense de 256 neurones avec activation ReLU. Pour prévenir le surapprentissage, une couche Dropout est intégrée avec un taux élevé de 0.7, supprimant aléatoirement une proportion significative de neurones durant l'entraînement.

Enfin, le modèle se termine par une couche de sortie dense composée d'un seul neurone activé par une fonction sigmoïde, adaptée à la tâche de classification binaire (cancer vs. normal). La structure complète du réseau est illustrée dans la Figure 10.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 5000, 32)          384
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 1000, 32)          0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 1000, 64)          18496
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 200, 64)           0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 200, 128)          57472
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 66, 128)           0
_________________________________________________________________
flatten (Flatten)            (None, 8448)              0
_________________________________________________________________
dense (Dense)                (None, 256)               2162944
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                16448
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
```

Figure 10 : Architecture du modèle CNN 1D conçu pour la classification binaire des profils d'expression génique

2.3. Entraînement du Modèle

L'entraînement du modèle convolutionnel a été réalisé à l'aide d'un GPU, dans le but de réduire le temps de calcul et d'optimiser l'efficacité du processus d'apprentissage.

Avant l'entraînement, les données ont été divisées en deux sous-ensembles : 80 % pour l'entraînement et 20 % pour le test, en utilisant la fonction train_test_split de la bibliothèque scikit-learn, comme illustré dans la Figure 11. Cette séparation permet de valider la capacité de généralisation du modèle sur des données non vues.

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

Figure 11 : Code pour partitionner le jeu de données

Afin d'harmoniser l'échelle des variables, une normalisation des données a été effectuée à l'aide du MinMaxScaler, qui met chaque caractéristique dans l'intervalle [0, 1]. Cette étape est cruciale pour garantir une convergence plus rapide et plus stable lors de l'entraînement des réseaux neuronaux.

Le modèle a été compilé avec l'optimiseur Adamax. La fonction de perte adoptée est la binary_crossentropy, appropriée pour les tâches de classification binaire. La précision (accuracy) a été utilisée comme métrique d'évaluation principale afin de suivre les performances pendant l'entraînement.

L'apprentissage a été conduit sur une durée de 40 époques, avec une taille de lot (batch size) fixée à 16, paramètre permettant un compromis entre vitesse d'exécution et stabilité des gradients.

Le script d'entraînement correspondant est présenté dans la Figure 12.

```
from tensorflow.keras.callbacks import EarlyStopping

# Define the early stopping callback
early_stopping = EarlyStopping(monitor='val_loss',
                               patience=20,
                               verbose=1, mode='min',
                               restore_best_weights=True)

# Fit the model with the early stopping callback
history = model.fit(
    X_train, y_train,
    epochs=100,
    batch_size=16,
    validation_data=(X_test, y_test),
    callbacks=[early_stopping, lr_scheduler]
)
```

Figure 12 : Code Python pour la compilation et l'entraînement du modèle

2.4. Évaluation des performances du modèle

L'évaluation du modèle a été menée à l'aide de plusieurs métriques de classification, intégrées à la bibliothèque scikit-learn. Les prédictions ont été obtenues sur les ensembles

d'entraînement et de test, puis converties en classes binaires à l'aide d'un seuil de 0.5, adapté à la fonction d'activation sigmoïde de la couche de sortie.

Les performances globales ont été mesurées à l'aide de la précision (accuracy), tandis qu'une analyse plus fine a été réalisée à l'aide du rapport de classification, qui inclut la précision, le rappel et le score F1 pour chaque classe (Figure 13).

```python
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt

# Get predictions
train_pred = model.predict(X_train)
test_pred = model.predict(X_test)

# Convert predictions to binary labels
train_pred_binary = (train_pred > 0.5).astype(int).flatten()
test_pred_binary = (test_pred > 0.5).astype(int).flatten()

# Convert true labels to binary
y_train_binary = y_train.flatten()
y_test_binary = y_test.flatten()

# Calculate accuracy
train_acc = accuracy_score(y_train_binary, train_pred_binary)
test_acc = accuracy_score(y_test_binary, test_pred_binary)

# Print accuracy
print("train-acc = " + str(train_acc))
print("test-acc = " + str(test_acc))
```

Figure 13 : Code utilisé pour l'évaluation des performances du modèle de prédiction.

En complément, une matrice de confusion normalisée a été générée afin de visualiser les proportions de prédictions correctes et incorrectes (Figure 14). Pour faciliter l'interprétation, cette matrice a été représentée sous forme de carte thermique à l'aide des bibliothèques Seaborn et Matplotlib. L'axe vertical a été inversé pour une correspondance visuelle plus intuitive avec les classes réelles. Cette approche permet de détecter les déséquilibres de classification ou les confusions fréquentes entre classes.

```python
# Compute confusion matrix
cm = confusion_matrix(y_test_binary, test_pred_binary)

# Reverse the order of rows in the confusion matrix
cm = np.flipud(cm)

# Normalize confusion matrix to get percentages
cm_percentage = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]

# Set up the plot with reversed y-axis
plt.figure(figsize=(8, 8))
yticklabels_reversed = list(reversed(label_binarizer.classes_))
sns.heatmap(cm_percentage, annot=True, fmt=".2%", cmap="Blues",
            xticklabels=label_binarizer.classes_, yticklabels=yticklabels_reversed)
plt.title('Normalized Confusion Matrix')
plt.xlabel('Predicted label')
plt.ylabel('True label')
plt.show()

# Print classification report
print(classification_report(y_test_binary, test_pred_binary,
                            target_names=label_binarizer.classes_))
```

Figure 14 : Code utilisé pour avoir la matrice de confusion.

CHAPITRE 3 : RÉSULTATS ET DISCUSSION

1. Résultats

L'apprentissage du modèle basé sur un réseau de neurones convolutif unidimensionnel (Conv1D) a donné des résultats encourageants pour la prédiction du cancer des poumons à partir des profils d'expression génique. La séparation des données en un ensemble d'entraînement (80 %) et un ensemble de test (20 %) a permis une évaluation rigoureuse et fiable de la capacité de généralisation du modèle. Ces performances valident l'intérêt de l'approche convolutionnelle pour l'extraction automatique de motifs biologiquement significatifs dans les données transcriptomiques.

1.1. Précision du Modèle

Le modèle a atteint une précision de 98,50 % sur les données d'entraînement et 99,00 % sur les données de test, témoignant d'une grande capacité de généralisation aux échantillons non vus. Ces résultats suggèrent que le modèle n'est pas sujet à un surapprentissage significatif et qu'il est apte à fournir des prédictions fiables sur de nouvelles données biologiques. Les résultats détaillés des prédictions sont illustrés dans la Figure 15.

```
train-acc = 0.9850746268656716
test-acc = 0.9900990099009901
```

Figure 15 : Résultats de la prédiction du modèle sur les ensembles d'entraînement et de test

1.2. Matrice de Confusion

La matrice de confusion, appliquée à l'ensemble de test, a permis une évaluation plus détaillée des performances du modèle au-delà de la simple précision. L'analyse de cette matrice révèle une forte proportion de vraies prédictions positives et négatives, indiquant que le modèle distingue efficacement les cas de cancer des échantillons normaux. Le nombre de faux positifs et faux négatifs reste faible, ce qui confirme la fiabilité du modèle pour la classification binaire. Ces résultats sont représentés visuellement dans la Figure 16.

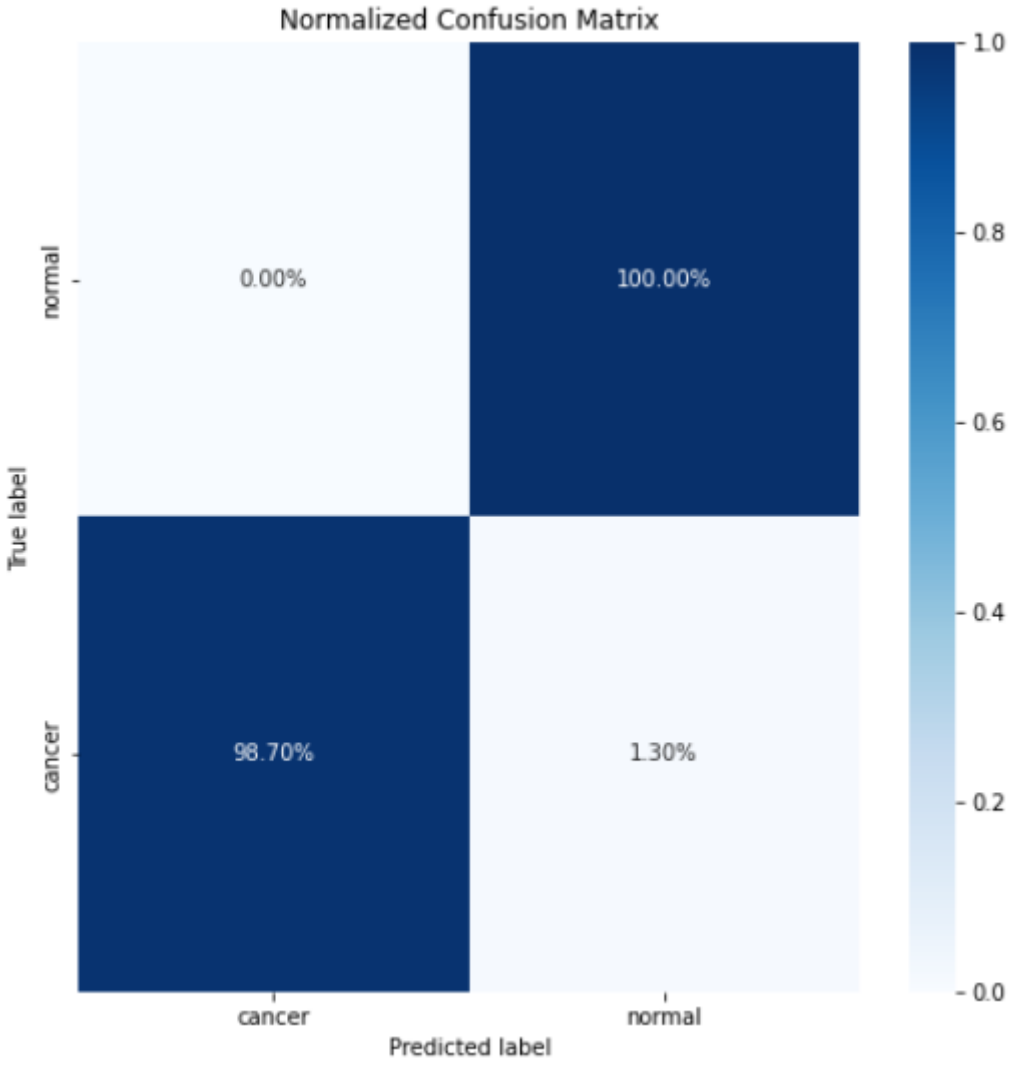

Figure 16 : Matrice de confusion normalisée pour l'évaluation du modèle sur l'ensemble de test

L'interprétation visuelle de la matrice de confusion normalisée (Figure 16) met en évidence les performances du modèle sur l'ensemble de test. Tous les échantillons appartenant à la classe "normal" ont été correctement classifiés, avec un taux de 100 % de vraies prédictions négatives. En revanche, pour la classe "cancer", 98,70 % des cas ont été correctement identifiés, tandis que 1,30 % ont été incorrectement prédits comme "normal", représentant ainsi les faux négatifs.

Fait notable : aucun faux positif n'a été observé, ce qui indique une spécificité parfaite du modèle pour la classe "normal". Cette distribution confirme une très haute sensibilité et spécificité, faisant de ce modèle un outil prometteur pour le dépistage du cancer à partir de données transcriptomiques.

1.3. Rapports de Classification

Le rapport de classification permet une évaluation plus détaillée des performances du modèle sur chaque classe. Comme le montre la Figure 17, la classe "cancer" a obtenu une précision parfaite de 100 %, un rappel de 99 % et un score F1 de 99 %, ce qui témoigne d'une excellente capacité du modèle à identifier correctement les cas positifs, tout en minimisant les faux négatifs.

Pour la classe "normal", bien que la précision soit légèrement inférieure (96 %), le modèle a atteint un rappel parfait de 100 %, indiquant que tous les échantillons sains ont été détectés correctement. Le score F1 pour cette classe atteint 98 %, démontrant un équilibre satisfaisant entre précision et rappel.

Globalement, le modèle atteint une précision moyenne (accuracy) de 99 %, avec des moyennes macro et pondérée (macro avg et weightedavg) également très élevées (≥ 98 %), ce qui confirme la robustesse du modèle et sa performance équilibrée sur l'ensemble des classes.

	precision	recall	f1-score	support
cancer	1.00	0.99	0.99	77
normal	0.96	1.00	0.98	24
accuracy			0.99	101
macro avg	0.98	0.99	0.99	101
weighted avg	0.99	0.99	0.99	101

Figure 17 : Rapport de classification indiquant les performances détaillées du modèle

2. Discussion

Les résultats obtenus démontrent l'efficacité du modèle Conv1D pour la prédiction du cancer à partir de données d'expression génique. Les taux de précision élevés observés tant sur l'ensemble d'entraînement que sur l'ensemble de test traduisent un modèle bien entraîné, capable de généraliser efficacement sur des données inédites. Cependant, malgré des performances globalement excellentes, la présence de faux positifs et de faux négatifs, même en faible proportion, invite à une analyse critique.

2.1. Interprétation des Résultats

L'obtention d'une précision de 99,00 % sur l'ensemble de test indique une capacité prédictive remarquable du modèle. Toutefois, les erreurs de classification — bien que rares — ont un impact significatif dans un contexte clinique. Les faux positifs, c'est-à-dire les cas où des échantillons sains sont prédits à tort comme cancéreux, peuvent entraîner des inquiétudes inutiles, voire des traitements non justifiés. À l'inverse, les faux négatifs, où des cas de cancer ne sont pas détectés, représentent un risque critique, pouvant conduire à un retard ou à une absence de prise en charge médicale.

Ces constats soulignent la nécessité d'améliorer la sensibilité et la spécificité du modèle, en particulier pour son application dans des environnements cliniques où les décisions doivent être fiables à 100 %. Des approches complémentaires pourraient être envisagées, telles que l'intégration de modèles hybrides, l'ajustement des seuils de classification, ou encore l'optimisation par validation croisée sur des ensembles de données plus diversifiés.

2.2. Perspectives d'amélioration et pistes futures

Bien que les résultats obtenus soient très encourageants, plusieurs pistes peuvent être explorées pour renforcer davantage la performance, la robustesse et la capacité de généralisation du modèle dans des contextes réels, notamment cliniques. Voici quelques axes d'amélioration potentiels :

- **Augmentation des données (Data Augmentation)** : L'intégration de nouvelles données d'expression génique, issues de bases de données complémentaires ou de cohortes cliniques variées, permettrait d'accroître la diversité biologique représentée dans l'ensemble d'apprentissage. Cette diversité est essentielle pour garantir une meilleure

généralisation du modèle face aux variations interindividuelles et aux différents sous-types de cancer. Par ailleurs, des techniques d'augmentation synthétique des données (par exemple via des algorithmes génératifs) pourraient également être envisagées, notamment pour rééquilibrer des classes sous-représentées.

- **Optimisation des Hyperparamètres** : L'efficacité d'un réseau de neurones dépend fortement de l'ajustement précis de ses hyperparamètres (nombre de couches, taille des noyaux, taux d'apprentissage, etc.). Une optimisation systématique, à l'aide de méthodes comme la recherche en grille (gridsearch), la recherche aléatoire, ou la recherche bayésienne, pourrait permettre d'identifier la combinaison optimale de paramètres pour améliorer les performances du modèle, tout en évitant le surapprentissage.
- **Validation croisée multi-stratégique** : La mise en place d'une validation croisée k-fold, combinée à une évaluation sur des cohortes indépendantes, renforcerait la fiabilité des résultats et aiderait à identifier d'éventuels biais liés aux données.
- **Déploiement clinique et interface utilisateur** : À terme, le modèle pourrait être intégré dans une plateforme logicielle conviviale destinée aux cliniciens ou chercheurs, avec un module d'interprétation des résultats, facilitant l'utilisation en situation réelle.

CONCLUSION

CONCLUSION :

Ce travail a permis de démontrer l'efficacité des réseaux de neurones convolutifs unidimensionnels (Conv1D) pour la classification binaire du cancer des poumons à partir de données d'expression génique. Grâce à une architecture optimisée et un entraînement rigoureux, le modèle a atteint une précision remarquable de 99,00 % sur l'ensemble de test, tout en conservant un taux d'erreur minimal.

L'analyse des métriques de performance, incluant la matrice de confusion et le rapport de classification, a confirmé la puissance du modèle, avec un excellent équilibre entre précision et rappel pour les deux classes. Malgré ces résultats très encourageants, la présence de quelques faux positifs et faux négatifs rappelle la nécessité de poursuivre les efforts en vue de renforcer la fiabilité du modèle, notamment dans une perspective d'application clinique.

Plusieurs axes d'amélioration ont été proposés, incluant l'enrichissement du jeu de données, l'optimisation fine des hyperparamètres, l'intégration de méthodes d'interprétabilité, et l'exploration de stratégies d'apprentissage ensembliste ou par transfert. À long terme, une validation sur des données réelles et hétérogènes ainsi que le déploiement dans un environnement applicatif pourraient permettre de faire de ce modèle un outil d'aide à la décision pour le dépistage précoce du cancer basé sur des signatures transcriptomiques.

RÉFÉRENCES BIBLIOGRAPHIQUES

REFERENCES BIBLIOGRAPHIQUES

Afify, H. M., Mohammed, K. K., & Hassanien, A. E. (2024). Leveraging hybrid 1D-CNN and RNN approach for classification of brain cancer gene expression. Complex & Intelligent Systems, 10(6), 7605-7617. https://doi.org/10.1007/s40747-024-01555-4

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Studying Gene Expression and Function. In Molecular Biology of the Cell. 4th edition. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26818/

Amicizia, D., Piazza, M. F., Marchini, F., Astengo, M., Grammatico, F., Battaglini, A., Schenone, I., Sticchi, C., Lavieri, R., Di Silverio, B., Andreoli, G. B., & Ansaldi, F. (2023). Systematic Review of Lung Cancer Screening : Advancements and Strategies for Implementation. Healthcare, 11(14), Article 14. https://doi.org/10.3390/healthcare11142085

ARCAGY-GINECO, D. B. P.-. (2025a, mai 17). La cancérisation et mutations génétiques. Infocancer. https://www.arcagy.org/infocancer/en-savoir-plus/le-cancer/qu-est-ce-que-le-cancer/canceristaion-et-mutation-genetique.html/?utm_source=chatgpt.com

ARCAGY-GINECO, D. B. P.-. (2025b, mai 23). L'intelligence artificielle (IA) en cancérologie. Infocancer. https://www.arcagy.org/infocancer/en-savoir-plus/canc-rologie-et-intelligence-artificielle.html/

Balan, A., Zhu, Q., Murray, J. C., Marrone, K. A., Scott, S. C., Feliciano, J. L., Hann, C. L., Ettinger, D. S., Smith, K. N., Forde, P. M., Brahmer, J. R., Levy, B. P., Elliott, A., VanderWalde, A., Oberley, M. J., Liu, S. V., Ma, P. C., Anders, R. A., & Anagnostou, V. (2023). Large-scale transcriptomic profiling of the tumor immune microenvironment in ALK+ lung cancer. Journal of Clinical Oncology. https://doi.org/10.1200/JCO.2023.41.16_suppl.9020

Basumallik, N., & Agarwal, M. (2025). Small Cell Lung Cancer. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK482458/

BioStudies. (s. d.). BioStudies < The European Bioinformatics Institute < EMBL-EBI. Consulté 23 mai 2025, à l'adresse https://www.ebi.ac.uk/biostudies/arrayexpress?utm_source=chatgpt.com

Bumgarner, R. (2013). DNA microarrays : Types, Applications and their future. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], 0 22, Unit-22.1. https://doi.org/10.1002/0471142727.mb2201s101

Chauhan, D. T. (2019, octobre 4). Gene Expression Microarray : Principle, Process, Advantages, Limitations and Applications. Genetic Education. https://geneticeducation.co.in/gene-expression-microarray/

Chen, D. T.-H., Hirst, J., Coupland, C. A. C., Liao, W., Baldwin, D. R., & Hippisley-Cox, J. (2025). Ethnic disparities in lung cancer incidence and differences in diagnostic characteristics : A population-based cohort study in England. The Lancet Regional Health – Europe, 48. https://doi.org/10.1016/j.lanepe.2024.101124

Clough, E., Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Lee, H., Zhang, N., Serova, N., Wagner, L., Zalunin, V., Kochergin, A., & Soboleva, A. (2024). NCBI GEO : Archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Research, 52(D1), D138-D144. https://doi.org/10.1093/nar/gkad965

Cognigni, V., Toscani, I., D'Agnelli, S., Pecci, F., Righi, L., Berardi, R., & Tiseo, M. (2025). Molecular heterogeneity of small cell lung cancer and new therapeutic possibilities : A narrative review of the literature. Translational Lung Cancer Research, 14(4). https://doi.org/10.21037/tlcr-24-755

Dépistage du cancer du poumon : La HAS recommande l'engagement d'un programme pilote. (s. d.). Haute Autorité de Santé. Consulté 17 mai 2025, à l'adresse https://www.has-sante.fr/jcms/p_3312901/en/depistage-du-cancer-du-poumon-la-has-recommande-l-engagement-d-un-programme-pilote

Dey, T. K., Mandal, S., & Mukherjee, S. (2022). Gene expression data classification using topology and machine learning models. BMC Bioinformatics, 22(10), 627. https://doi.org/10.1186/s12859-022-04704-z

Diagnostic du cancer du poumon – Centres interdisciplinaires d'oncologie. (s. d.). Consulté 17 mai 2025, à l'adresse https://centrescancer.chuv.ch/etapediagnostictrait/diagnostic-du-cancer-du-poumon/?utm_source=chatgpt.com

Frequently Asked Questions—GEO - NCBI. (s. d.). Consulté 23 mai 2025, à l'adresse https://www.ncbi.nlm.nih.gov/geo/info/faq.html?utm_source=chatgpt.com#what

Fu, K., Xie, F., Wang, F., & Fu, L. (2022). Therapeutic strategies for EGFR-mutated non-small cell lung cancer patients with osimertinib resistance. Journal of Hematology & Oncology, 15(1), 173. https://doi.org/10.1186/s13045-022-01391-4

Fu, Y.-C., Liang, S.-B., Luo, M., & Wang, X.-P. (2025). Intratumoral heterogeneity and drug resistance in cancer. Cancer Cell International, 25(1), 103. https://doi.org/10.1186/s12935-025-03734-w

Gabriel, A. A. G., Atkins, J. R., Penha, R. C. C., Smith-Byrne, K., Gaborieau, V., Voegele, C., Abedi-Ardekani, B., Milojevic, M., Olaso, R., Meyer, V., Boland, A., Deleuze, J. F., Zaridze, D., Mukeriya, A., Swiatkowska, B., Janout, V., Schejbalová, M., Mates, D., Stojšić, J., … McKay, J. D. (2022). Genetic Analysis of Lung Cancer and the Germline Impact on Somatic Mutation Burden. Journal of the National Cancer Institute, 114(8), 1159-1166. https://doi.org/10.1093/jnci/djac087

Gene expression profiling in cancer. (2025). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Gene_expression_profiling_in_cancer&oldid=1292380505

GeneChip™ Human Genome U133 Plus 2.0 Array. (s. d.). Consulté 23 mai 2025, à l'adresse https://www.thermofisher.com/order/catalog/product/900467

Gonzalez, L., Alcaraz, A., Gabay, C., Castro, M., Vigo, S., Carinci, E., & Augustovski, F. (2023). Health-related Quality of life, Financial Toxicity, Productivity Loss and Catastrophic

Health Expenditures After Lung Cancer Diagnosis in Argentina (arXiv:2312.16710). arXiv. https://doi.org/10.48550/arXiv.2312.16710

Gunavathi, C., Sivasubramanian, K., Keerthika, P., & Paramasivam, C. (2021). A review on convolutional neural network based deep learning methods in gene expression data for disease diagnosis. Materials Today: Proceedings, 45, 2282-2285. https://doi.org/10.1016/j.matpr.2020.10.263

How to transform lung cancer outcomes in LMICs. (2024, novembre 29). World Economic Forum. https://www.weforum.org/stories/2024/11/how-to-transform-lung-cancer-outcomes-in-lmics/

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 15(1), 41-51.

Hwang, H., Jeon, H., Yeo, N., & Baek, D. (2024). Big data and deep learning for RNA biology. Experimental & Molecular Medicine, 56(6), 1293-1321. https://doi.org/10.1038/s12276-024-01243-w

Imyanitov, E. N., Preobrazhenskaya, E. V., & Orlov, S. V. (2024). Current status of molecular diagnostics for lung cancer. Exploration of Targeted Anti-Tumor Therapy, 5(3), Article 3. https://doi.org/10.37349/etat.2024.00244

Introduction aux réseaux neuronaux convolutifs (CNN). (2025). https://fr.mathworks.com/discovery/convolutional-neural-network.html

Introduction to Convolution Neural Network. (2025). GeeksforGeeks. https://www.geeksforgeeks.org/introduction-convolution-neural-network/

Islami, F., Torre, L. A., & Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. Translational Lung Cancer Research, 4(4). https://doi.org/10.3978/j.issn.2218-6751.2015.08.04

Jaluria, P., Konstantopoulos, K., Betenbaugh, M., & Shiloach, J. (2007). A perspective on microarrays : Current applications, pitfalls, and potential uses. Microbial Cell Factories, 6, 4. https://doi.org/10.1186/1475-2859-6-4

Jelic, S. & MD. (s. d.). Small Cell vs. Non-Small Cell Lung Cancer. Verywell Health. Consulté 17 mai 2025, à l'adresse https://www.verywellhealth.com/small-cell-vs-non-small-cell-lung-cancer-5208050

Jha, A., Quesnel-Vallières, M., Wang, D., Thomas-Tikhonenko, A., Lynch, K. W., & Barash, Y. (2022). Identifying common transcriptome signatures of cancer by interpreting deep learning models. Genome Biology, 23, 117. https://doi.org/10.1186/s13059-022-02681-3

Kaissar, A., Nassif, A. B., Soudan, B., & Injadat, M. (2025). Enhancing CNN-based network intrusion detection through hyperparameter optimization. Intelligent Systems with Applications, 26, 200528. https://doi.org/10.1016/j.iswa.2025.200528

Kakati, T., Bhattacharyya, D. K., Kalita, J. K., & Norden-Krichmar, T. M. (2022). DEGnext : Classification of differentially expressed genes from RNA-seq data using a convolutional

neural network with transfer learning. BMC Bioinformatics, 23(1), 17. https://doi.org/10.1186/s12859-021-04527-4

Karimzadeh, M., Momen-Roknabadi, A., Cavazos, T. B., Fang, Y., Chen, N.-C., Multhaup, M., Yen, J., Ku, J., Wang, J., Zhao, X., Murzynowski, P., Wang, K., Hanna, R., Huang, A., Corti, D., Nguyen, D., Lam, T., Kilinc, S., Arensdorf, P., … Goodarzi, H. (2024). Deep generative AI models analyzing circulating orphan non-coding RNAs enable detection of early-stage lung cancer. Nature Communications, 15(1), 10090. https://doi.org/10.1038/s41467-024-53851-9

Kim, W.-J., Choi, B. R., Noh, J. J., Lee, Y.-Y., Kim, T.-J., Lee, J.-W., Kim, B.-G., & Choi, C. H. (2024). Comparison of RNA-Seq and microarray in the prediction of protein expression and survival prediction. Frontiers in Genetics, 15. https://doi.org/10.3389/fgene.2024.1342021

La bioinformatique connaît une innovation significative grâce à l'IA et à l'apprentissage automatique | HackerNoon. (s. d.). Consulté 23 mai 2025, à l'adresse https://hackernoon.com/lang/fr/la-bioinformatique-connaît-une-innovation-importante-grâce-à-l'IA-et-à-l'apprentissage-automatique

Li, C., Lei, S., Ding, L., Xu, Y., Wu, X., Wang, H., Zhang, Z., Gao, T., Zhang, Y., & Li, L. (2023). Global burden and trends of lung cancer incidence and mortality. Chinese Medical Journal, 136(13), 1583-1590. https://doi.org/10.1097/CM9.0000000000002529

Li, C., Nguyen, T. T., Li, J.-R., Song, X., Fujimoto, J., Little, L., Gumb, C., Chow, C.-W. B., Wistuba, I. I., Futreal, A. P., Zhang, J., Hubert, S. M., Heymach, J. V., Wu, J., Amos, C. I., Zhang, J., & Cheng, C. (2024). Multiregional transcriptomic profiling provides improved prognostic insight in localized non-small cell lung cancer. Npj Precision Oncology, 8(1), 1-14. https://doi.org/10.1038/s41698-024-00680-0

L'IA transforme le combat contre le cancer avec des traitements personnalisés. (2024, avril 4). https://www.ictjournal.ch/articles/2024-04-04/lia-transforme-le-combat-contre-le-cancer-avec-des-traitements-personnalises

Liu, S., & Yao, W. (2022). Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. BMC Bioinformatics, 23(1), 175. https://doi.org/10.1186/s12859-022-04689-9

Lung cancer. (2025). https://www.who.int/news-room/fact-sheets/detail/lung-cancer

Lung cancer screening. (2025). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Lung_cancer_screening&oldid=1287287359

Lung Cancer Treatment & Survival Rate | City of Hope. (2025, février 7). https://www.cityofhope.org/clinical-program/lung-cancer/treatments-survival

Lung-MAP Clinical Trial—NCI (nciglobal,ncienterprise). (2014, juin 16). [cgvArticle]. https://www.cancer.gov/types/lung/research/lung-map

Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology, 9(1), 34. https://doi.org/10.1186/1741-7007-9-34

Mazzone, P. J., & Lam, L. (2022). Evaluating the Patient With a Pulmonary Nodule : A Review. JAMA, 327(3), 264. https://doi.org/10.1001/jama.2021.24287

Microarray | Learn Science at Scitable. (s. d.). Consulté 23 mai 2025, à l'adresse https://www.nature.com/scitable/definition/microarray-202/?utm_source=chatgpt.com

Mohamed, T. I. A., Ezugwu, A. E., Fonou-Dombeu, J. V., Ikotun, A. M., & Mohammed, M. (2023). A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data. Scientific Reports, 13(1), 14644. https://doi.org/10.1038/s41598-023-41731-z

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., & Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Scientific Reports, 11(1), 15626. https://doi.org/10.1038/s41598-021-95128-x

Nwagbara, U. I., Ginindza, T. G., & Hlongwana, K. W. (2020). Health systems influence on the pathways of care for lung cancer in low- and middle-income countries : A scoping review. Globalization and Health, 16(1), 23. https://doi.org/10.1186/s12992-020-00553-8

Parisapogu, S. A. B., Annavarapu, C. S. R., & Elloumi, M. (2021). 1-Dimensional Convolution Neural Network Classification Technique for Gene Expression Data. In M. Elloumi (Éd.), Deep Learning for Biomedical Data Analysis : Techniques, Approaches, and Applications (p. 3-26). Springer International Publishing. https://doi.org/10.1007/978-3-030-71676-9_1

Patharia, P., Sethy, P. K., & Nanthaamornphong, A. (2024). Advancements and Challenges in the Image-Based Diagnosis of Lung and Colon Cancer : A Comprehensive Review. Cancer Informatics, 23, 11769351241290608. https://doi.org/10.1177/11769351241290608

(PDF) Imagerie radiologique et TEP Scanner dans les cancers du poumon. (2024). ResearchGate. https://doi.org/10.1016/S0221-0363(08)89016-6

Pleasance, E., Bohm, A., Williamson, L. M., Nelson, J. M. T., Shen, Y., Bonakdar, M., Titmuss, E., Csizmok, V., Wee, K., Hosseinzadeh, S., Grisdale, C. J., Reisle, C., Taylor, G. A., Lewis, E., Jones, M. R., Bleile, D., Sadeghi, S., Zhang, W., Davies, A., … Laskin, J. (2022). Whole-genome and transcriptome analysis enhances precision cancer treatment options. Annals of Oncology, 33(9), 939-949. https://doi.org/10.1016/j.annonc.2022.05.522

Qu'est-ce que l'intelligence artificielle (IA) et pourquoi est-elle importante | NetApp. (s. d.). Consulté 23 mai 2025, à l'adresse https://www.netapp.com/fr/artificial-intelligence/what-is-artificial-intelligence/

Rapport de 2024 de l'Institut national de santé publique : Hausse des cas de cancer à Alger en 2021. (2024). El Watan. https://elwatan-dz.com/rapport-de-2024-de-linstitut-national-de-sante-publique-hausse-des-cas-de-cancer-a-alger-en-2021

Robert, J. (2020a, septembre 28). Deep Learning ou Apprentissage Profond : Qu'est-ce que c'est ? DataScientest. https://datascientest.com/deep-learning-definition

Robert, J. (2020b, novembre 18). Machine Learning : Définition, fonctionnement, utilisations. DataScientest. https://datascientest.com/machine-learning-tout-savoir

Saggi, M. K., Bhatia, A. S., Isaiah, M., Gowher, H., & Kais, S. (2024). Multi-Omic and Quantum Machine Learning Integration for Lung Subtypes Classification (arXiv:2410.02085). arXiv. https://doi.org/10.48550/arXiv.2410.02085

S'attaquer à l'impact du cancer sur la santé, l'économie et la société : France. (2024, novembre 20). OCDE. https://www.oecd.org/fr/publications/s-attaquer-a-l-impact-du-cancer-sur-la-sante-l-economie-et-la-societe_0f779a1e-fr/france_4609e421-fr.html

Singhal, S., Miller, D., Ramalingam, S., & Sun, S.-Y. (2008). Gene Expression Profiling of Non-Small Cell Lung Cancer. Lung cancer (Amsterdam, Netherlands), 60(3), 313-324. https://doi.org/10.1016/j.lungcan.2008.03.007

Sorscher, S., LoPiccolo, J., Heald, B., Chen, E., Bristow, S. L., Michalski, S. T., Nielsen, S. M., Lacoste, A., Keyder, E., Lee, H., Nussbaum, R. L., Martins, R., & Esplin, E. D. (2023). Rate of Pathogenic Germline Variants in Patients With Lung Cancer. JCO Precision Oncology, 7, e2300190. https://doi.org/10.1200/PO.23.00190

Sun, B., & Chen, L. (2023). Interpretable deep learning for improving cancer patient survival based on personal transcriptomes. Scientific Reports, 13(1), 11344. https://doi.org/10.1038/s41598-023-38429-7

Šutić, M., Dmitrović, B., Jakovčević, A., Džubur, F., Oršolić, N., Debeljak, Ž., Försti, A., Seiwerth, S., Brčić, L., Madzarac, G., Samaržija, M., Jakopović, M., & Knežević, J. (2024). Transcriptomic Profiling for Prognostic Biomarkers in Early-Stage Squamous Cell Lung Cancer (SqCLC). Cancers, 16(4), 720. https://doi.org/10.3390/cancers16040720

Tarca, A. L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. American journal of obstetrics and gynecology, 195(2), 373-388. https://doi.org/10.1016/j.ajog.2006.07.001

The Cancer Genome Atlas. (2025). In Wikipedia. https://en.wikipedia.org/w/index.php?title=The_Cancer_Genome_Atlas&oldid=1278608823

Traitement et pronostic du cancer du poumon non à petites cellules. (s. d.). Consulté 17 mai 2025, à l'adresse https://www.jnjmedicalcloud.ch/fr-ch/therapeutic-area/onkologie/lungenkrebs?utm_source=chatgpt.com

Traitement personnalisé du cancer avec l'IA : s'orienter dans le paysage génomique. (s. d.). Consulté 23 mai 2025, à l'adresse https://www.apollohospitals.com/fr/health-library/personalized-cancer-treatment-navigating-the-genomic-landscape-with-apollo-precision-oncology-centres-ai-solutions?utm_source=chatgpt.com

Transcriptomique. (2025). In Wikipédia. https://fr.wikipedia.org/w/index.php?title=Transcriptomique&oldid=225801946

tropicale, A. S. (2025). 85% des cancers du poumon liés au tabagisme. https://www.santemaghreb.com/actus.asp?id=25680&utm_source=chatgpt.com

Types of Lung Cancer | LUNGevity Foundation. (s. d.). Consulté 17 mai 2025, à l'adresse https://www.lungevity.org/lung-cancer-basics/types-of-lung-cancer?utm_source=chatgpt.com

Using TCGA - NCI (nciglobal,ncienterprise). (2019, mars 6). [cgvArticle]. https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data

Vekris, A., & Robert, J. (2005). Prédiction de la réponse aux anticancéreux par analyse du transcriptome. Les nouvelles voies de la pharmacogénomie. Oncologie, 7(1), 17-23. https://doi.org/10.1007/s10269-005-0147-7

Wang, P., Ng, R., Lam, S., & Lockwood, W. W. (2024). Uncovering molecular features driving lung adenocarcinoma heterogeneity in patients who formerly smoked. Journal of Translational Medicine, 22(1), 634. https://doi.org/10.1186/s12967-024-05437-8

Wen, J., Fu, J., Zhang, W., & Guo, M. (2011). Genetic and epigenetic changes in lung carcinoma and their clinical implications. Modern Pathology, 24(7), 932-943. https://doi.org/10.1038/modpathol.2011.46

Wilkinson, A. N., & Lam, S. (2021). ABC du dépistage du cancer du poumon. Canadian Family Physician, 67(11), 823-829. https://doi.org/10.46747/cfp.6711823

Word Health Organization. (2025). Cancer du poumon. https://www.who.int/fr/news-room/fact-sheets/detail/lung-cancer

Wu, J., & Lin, Z. (2022). Non-Small Cell Lung Cancer Targeted Therapy : Drugs and Mechanisms of Drug Resistance. International Journal of Molecular Sciences, 23(23), 15056. https://doi.org/10.3390/ijms232315056

Yoh, K. (2019). Novel targeted therapy beyond EGFR and ALK : ROS1, BRAF, RET and MET. Annals of Oncology, 30, vi35. https://doi.org/10.1093/annonc/mdz328

Z. Melissa. (2024, septembre 19). Pollution de l'air en Algérie : 50 personnes décédées en une année. https://www.algerie360.com/pollution-de-lair-en-algerie-50-personnes-decedees-en-une-annee/

Zarinshenas, R., Amini, A., Mambetsariev, I., Abuali, T., Fricke, J., Ladbury, C., & Salgia, R. (2023). Assessment of Barriers and Challenges to Screening, Diagnosis, and Biomarker Testing in Early-Stage Lung Cancer. Cancers, 15(5), Article 5. https://doi.org/10.3390/cancers15051595

Zengin, T., & Önal-Süzek, T. (2020). Analysis of Genomic and Transcriptomic Variations as Prognostic Signature for Lung Adenocarcinoma. BMC Bioinformatics, 21(S14), 368. https://doi.org/10.1186/s12859-020-03691-3

Zhou, N., Xu, Y., Huang, Y., Ye, G., Luo, L., & Song, Z. (2025). Comprehensive genomic profiling of Chinese lung cancer characterizes germline-somatic mutation interactions influencing cancer risk. Journal of Translational Medicine, 23(1), 199. https://doi.org/10.1186/s12967-025-06096-z

Soutenu le :	Présenté par :
25/06/2025	SAKLOUL Malak

Thème :

Prédiction du cancer des poumons à partir des données d'expression des gènes basée sur le Deeplearning

Mémoire Présenté en vue de l'obtention du Diplôme de Master en :

Bioinformatique

Domaine : Science de la nature et la vie

Département de Biologie Appliquée

L'identification précoce du cancer à partir des données transcriptomiques constitue un enjeu majeur en bioinformatique et en médecine de précision. Dans cette étude, nous avons exploré l'efficacité des réseaux de neurones convolutifs unidimensionnels (Conv1D) pour la classification binaire des profils d'expression génique, en distinguant les échantillons cancéreux des échantillons normaux. Les données utilisées proviennent d'un fichier annoté nommé LUNG_cancer_labeled_5000.csv, contenant des profils transcriptomiques préalablement étiquetés.

Le modèle développé a été entraîné sur 80 % des données et testé sur les 20 % restantes. Il repose sur une architecture composée de couches Conv1D avec activation ReLU, couches de MaxPooling, suivies d'une couche dense et d'un Dropout pour limiter le surapprentissage. Les performances ont été évaluées à l'aide de métriques classiques : précision, rappel, F1-score, matrice de confusion et rapport de classification.

Les résultats obtenus sont très prometteurs, avec une précision de 98,8 % sur l'ensemble de test, une matrice de confusion montrant une excellente capacité de détection des deux classes, et un déséquilibre minimal entre faux positifs et faux négatifs. Ces performances soulignent le potentiel des Conv1D pour des applications biomédicales concrètes.

Des perspectives d'amélioration ont été proposées, notamment l'augmentation des données, l'optimisation des hyperparamètres, l'intégration d'outils d'explicabilité et le recours à des approches d'apprentissage ensembliste. Ces pistes ouvrent la voie vers un déploiement futur dans des environnements cliniques de dépistage assisté par intelligence artificielle.

Mots clés : Apprentissage profond, Conv1D, Expression génique, Classification binaire, Cancer du poumon, Transcriptomique, Bioinformatique, Réseaux neuronaux, Intelligence artificielle, Dépistage précoce.

Jury d'évaluation :

Président du jury : Dr. DAAS Mohamed Skander

Encadreur : Pr. BELIL Ines

Examinatrice : Dr. AMINE KHOUDJA ihsein Rokia