

Université d'Aïn Témouchent - Belhadj Bouchaïb

Faculté des Sciences et de Technologie

Département de Mathématiques et d'Informatique



Projet de Fin d'Études

Pour l'obtention du diplôme de Master en Informatique



Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Cybersécurité et Intelligence Artificielle (CYSIA)

THÈME

Modèle intelligent pour le diagnostic médical Application aux données multimodales

PRÉSENTÉ PAR :

- Mr. Bouhacida Ibrahim
- Mr. Saïm Sofiane

DEVANT LE JURY COMPOSÉ DE :

Nom et Prénom	Grade	Université	Rôle
Dr. Belgrana .F/Z	M.C.A	UAT.B.B	Présidente
Dr.Mohammed Hakim BEN-DIABDALLAH	M.C.A	UAT.B.B	Examineur
Dr. BENOMAR MOHAM-MED LAMINE	M.C.A	UAT.B.B	Encadrant



ANNÉE UNIVERSITAIRE 2025/2026

Dédicace

À mes parents,
pour leur sacrifice et leur amour inconditionnel.

À mon épouse,
mon pilier, ma force au quotidien.

À mes deux merveilleuses filles,
pour qui chaque effort a du sens.

À mes quatre frères,
pour leur soutien indéfectible, leur complicité et leur présence à mes côtés.

À mon cher neveu,
qui apporte tant de joie à notre famille.

À mon encadreur,
pour sa confiance et ses précieux conseils.

À tous ceux qui ont marqué mon parcours personnel et professionnel,
qu'ils reçoivent ici le témoignage de ma sincère gratitude.

Ibrahim

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mon encadrant de mémoire, M. **Benomar Mohammed Lamine**, pour sa disponibilité, ses précieux conseils et son soutien constant tout au long de la réalisation de ce travail. Ses remarques avisées et sa rigueur scientifique m'ont guidé à chaque étape de ce projet.

Mes sincères remerciements s'adressent également aux membres du jury qui ont accepté d'évaluer ce travail et d'honorer de leur présence la soutenance de ce mémoire.

Je remercie chaleureusement l'ensemble du corps professoral et administratif du département de **Mathématiques et d'Informatique** pour la qualité de la formation qui m'a été dispensée tout au long de mon parcours universitaire.

Un grand merci à mes camarades de promotion pour les moments d'entraide, les discussions stimulantes et la bonne humeur partagée durant cette année de Master.

Enfin, je ne saurais clore ces remerciements sans exprimer ma gratitude infinie à ma famille et à mes proches. Leur soutien moral, leur patience et leurs encouragements inconditionnels ont été mes plus grands moteurs pour mener à bien mes études.

Table des matières

Introduction générale	8
Chapitre 1 : Données et images médicales	10
1.1 Introduction	11
1.2 Données médicales	11
1.3 Diagnostic médical	12
1.4 Image numérique	13
1.4.1 Formats des images numériques	19
1.5 Imagerie médicale	19
1.6 Prétraitement et analyse des images médicales	21
1.7 Compte rendu médical (Rapport)	22
1.8 Conclusion	23
Chapitre 2 : Classification multimodale des données médicales	25
2.1 Introduction	26
2.2 IA dans le domaine médical	26
2.3 Machine learning vs deep learning	27
2.4 Classification des données médicales	28
2.4.1 Métriques d'évaluation des performances	29
2.4.2 Classification d'images médicales par CNN	30
2.4.3 Extraction des caractéristiques visuelles avec CNN	32
2.4.4 Traitement automatique du langage (texte) médical	33
2.4.5 Transformers et BERT pour les rapports médicaux	35
2.5 Approche multimodal en diagnostic médical	37
2.6 Les types de fusion	38
2.7 État de l'art des approches multimodal (CNN, BERT)	39
2.8 Conclusion	41
Chapitre 3 : Expérimentations, résultats et discussion	43
3.1 Introduction	44
3.2 Description et préparation des données	44
3.2.1 Jeu de données Indiana University Chest X-ray	44
3.2.2 Prétraitement des rapports textuels	46
3.2.3 Regroupement et nettoyage des labels	46

3.2.4	Fusion des données textuelles et images radiographiques	48
3.3	Méthodologie : pipeline multimodal proposé	49
3.3.1	Explication de l'architecture proposée	50
3.4	Résultats expérimentaux	54
3.4.1	Courbes d'apprentissage (loss)	54
3.4.2	Performances globales	55
3.4.3	Matrices de confusion par label	56
3.4.4	Courbes ROC et aires sous la courbe (AUC)	58
3.4.5	Analyse par groupe diagnostique	60
3.5	Discussion	61
3.5.1	Interprétation des performances	61
3.5.2	Comparaison contextuelle avec d'autres stratégies de fusion	62
3.5.3	Analyse comparative des stratégies de fine-tuning	63
3.5.4	Limites de l'approche	67
3.5.5	Pistes d'amélioration	67
3.6	Interface de démonstration : Gradio	67
3.6.1	Architecture logicielle et flux de traitement	68
3.6.2	Illustration de l'interface sur un cas réel	68
3.7	Conclusion	69
	Conclusion générale	70
	Synthèse de la littérature	71
	Bibliographie	79

Table des figures

1.1	Représentation d'une image numérique sous forme de matrice de pixels . . .	14
1.2	Dimension d'une image numérique	15
1.3	Illustration de la résolution spatiale	15
1.4	Représentation des vecteurs gradients selon les variations de luminance . . .	16
1.5	Effet du contraste sur la visibilité des détails	16
1.6	Exemple de contours	17
1.7	Segmentation des régions et types pulmonaires	17
1.8	Illustration de l'impact du bruit et des filtres de restauration	18
1.9	Illustration de la notion de texture	18
1.10	Image en niveaux de gris et histogramme correspondant	19
1.11	Exemple de radiographie thoracique	20
1.12	Exemple de compte rendu radiologique réel	23
2.1	Comparaison entre le machine learning et le deep learning	28
2.2	Les trois types de classification en diagnostic médical.	29
2.3	Architecture générale d'un CNN	31
2.4	Techniques de représentation textuelle en NLP médical	34
2.5	Architecture du modèle Transformer	35
2.6	Processus de pré-entraînement et de fine-tuning de BERT	36
2.7	Principales sources de données multimodales dans le domaine biomédical . .	37
2.8	Comparaison des performances des cinq stratégies de fusion multimodale . .	39
3.1	Extrait du fichier indiana_reports.csv	45
3.2	Extrait du fichier indiana_projections.csv	45
3.3	Pipeline du modèle multimodal de diagnostic thoracique.	50
3.4	Architecture générale du réseau DenseNet121	51
3.5	Architecture de BioClinicalBERT	52
3.6	Courbes d'apprentissage de la perte (BCEWithLogitsLoss)	55
3.7	Matrices de confusion par label	57
3.8	Valeurs d'AUC par classe sur l'ensemble de test	59
3.9	Courbes ROC par classe sur l'ensemble de test	59
3.10	Démonstration de l'interface Gradio lors de l'évaluation d'un cas multimodal.	69

Liste des tableaux

1.1	Types d'images numériques	13
3.1	Nettoyage des comptes rendus : concaténation des sections <i>findings</i> et <i>impression</i>	46
3.2	Nettoyage et regroupement des labels	46
3.3	Regroupement des classes pour la classification multimodale	47
3.4	Réduction du nombre de patients après fusion avec les images radiographiques	49
3.5	Évolution de la distribution des classes avant et après fusion	49
3.6	Métriques d'évaluation globales du modèle multimodal sur le test set	56
3.7	Métriques de performance par classe de pathologie (TP, FP, FN, TN)	58
3.8	Métriques détaillées par groupe de pathologies sur le test set	60
3.9	Comparaison des stratégies de fusion sémantique	62
3.10	Analyse incrémentale des stratégies de fine-tuning de l'architecture unifiée .	66

Introduction générale

Introduction générale

L'intelligence artificielle occupe aujourd'hui une place importante dans le domaine de la santé, notamment dans l'analyse des données médicales et l'assistance au diagnostic [1]. Grâce aux progrès de l'apprentissage automatique et, plus particulièrement, de l'apprentissage profond, il est désormais possible de développer des systèmes capables d'extraire automatiquement des informations pertinentes à partir de données complexes telles que les images médicales, les rapports cliniques ou les dossiers patients[2]. Les réseaux de neurones convolutionnels, ou CNN, ont démontré leur efficacité dans l'analyse et la classification des images médicales grâce à leur capacité à extraire automatiquement des caractéristiques visuelles pertinentes [2] [3]. Parallèlement, les architectures de type Transformer ont marqué une avancée importante dans le traitement du langage naturel grâce au mécanisme d'attention [4]. Parmi ces modèles, BERT a obtenu des résultats remarquables dans la compréhension et la représentation des données textuelles [5], y compris dans le contexte des textes cliniques et médicaux [6].

Dans la pratique médicale, le diagnostic repose rarement sur une seule source d'information. Le médecin confronte généralement plusieurs types de données, comme les images radiologiques, les symptômes, les antécédents du patient et les comptes rendus médicaux, afin de prendre une décision plus fiable [7]. Cependant, de nombreux modèles d'intelligence artificielle restent encore limités à une approche unimodale, en exploitant uniquement l'image ou uniquement le texte. Cette limitation peut réduire leur capacité à représenter toute la richesse de l'information médicale disponible [7] [8].

Dans ce contexte, l'apprentissage multimodal constitue une approche prometteuse, car il permet de combiner plusieurs sources d'information complémentaires afin d'améliorer la représentation des données et les performances des modèles [8]. L'association d'un CNN et de BERT permet ainsi de combiner les caractéristiques visuelles extraites des images médicales avec les informations textuelles issues des rapports médicaux. Des travaux récents ont montré que l'intégration de données d'imagerie et de données cliniques peut améliorer les performances des modèles par rapport aux approches unimodales [7].

Ce projet de fin d'études s'inscrit dans cette perspective et propose une approche de classification multimodale basée sur la combinaison d'un modèle CNN pour l'analyse des images médicales et d'un modèle BERT pour le traitement des données textuelles. L'objectif principal est de montrer que la fusion de ces deux modalités peut améliorer les performances de classification, tout en offrant un système d'aide à la décision médicale plus robuste et plus proche du raisonnement clinique.

Ce mémoire est organisé en trois chapitres. Le premier chapitre présente les notions générales liées aux données médicales, à l'imagerie médicale et au traitement automatique du langage (texte) médical. Le deuxième chapitre est consacré aux méthodes d'apprentissage

profond utilisées dans ce travail, notamment les CNN, les modèles BERT et les stratégies de fusion multimodale. Un état de l'art de ces différentes approches est également présenté. Le troisième chapitre présente la partie expérimentale, les outils utilisés, les résultats obtenus ainsi que leur discussion. Enfin, le mémoire se termine par une conclusion générale et quelques perspectives pour des travaux futurs.

Chapitre 1 : Données et images médicales

1.1 Introduction

L'évolution des technologies numériques a profondément transformé le domaine de la santé, où les données médicales occupent désormais une place centrale dans le diagnostic, le suivi thérapeutique et la prise de décision clinique. Ces données sont de nature variée et comprennent notamment les informations administratives, les données biologiques, les images médicales ainsi que les comptes rendus rédigés par les spécialistes. L'exploitation efficace de ces différentes sources d'information constitue aujourd'hui un enjeu majeur pour le développement des systèmes d'aide au diagnostic fondés sur l'intelligence artificielle [9, 2].

Parmi ces données, l'imagerie médicale représente l'une des sources d'information les plus importantes. Grâce à des modalités telles que la radiographie, la tomodensitométrie (scanner), l'imagerie par résonance magnétique (IRM) ou encore l'échographie, elle permet de visualiser les structures anatomiques et de détecter de nombreuses anomalies pathologiques. Toutefois, l'interprétation des images est généralement complétée par des rapports médicaux décrivant les observations du radiologue et leur signification clinique [10].

Avant d'aborder les méthodes d'apprentissage profond et les approches multimodales utilisées dans ce mémoire, il est indispensable de présenter les concepts fondamentaux relatifs aux données médicales et aux images numériques. Une bonne compréhension de ces notions permet de mieux appréhender les différentes étapes de préparation, d'analyse et d'exploitation des données utilisées dans les systèmes modernes d'aide au diagnostic.

Ce chapitre présente tout d'abord les principales catégories de données médicales ainsi que le processus de diagnostic médical. Il introduit ensuite les notions fondamentales relatives aux images numériques, leurs caractéristiques et leurs formats. Les principales modalités d'imagerie médicale sont ensuite décrites, avant d'aborder le prétraitement des images ainsi que les comptes rendus médicaux, qui constituent la seconde modalité exploitée dans ce travail.

1.2 Données médicales

Les données médicales désignent l'ensemble des informations recueillies tout au long de la vie d'un patient dans le cadre de son suivi, de son diagnostic et de sa prise en charge médicale. Elles sont multiples et de natures diverses. Elles peuvent comprendre des données administratives, telles que l'identité, l'âge et le sexe du patient, ainsi que des données cliniques relatives à son état de santé. Ces dernières incluent les antécédents médicaux, les symptômes rapportés, les observations issues de l'examen clinique, les résultats d'analyses biologiques, les paramètres physiologiques et les examens complémentaires [7].

Dans le cadre du diagnostic médical, deux types de données occupent une place particulièrement importante : l'imagerie médicale et les rapports de radiologie. L'imagerie médicale fournit une information visuelle permettant d'observer les structures anatomiques ou les anomalies pathologiques [11]. Les rapports de radiologie représentent une information textuelle rédigée par le spécialiste afin de décrire les observations, proposer une interprétation médicale et orienter la prise en charge du patient [12].

Ainsi, selon les données disponibles et l'objectif clinique, le diagnostic peut être basé sur l'analyse de l'imagerie médicale, sur l'interprétation des rapports de radiologie, ou sur la combinaison de ces deux sources d'information. Cette complémentarité entre l'image et le texte constitue une base importante pour les approches multimodales en intelligence artificielle médicale [7] [12].

Parmi les différentes catégories de données médicales, l'imagerie médicale occupe une place centrale. Elle regroupe un ensemble de techniques permettant d'obtenir des représentations visuelles du corps humain à des fins diagnostiques ou thérapeutiques. Selon la modalité utilisée, l'image peut être obtenue à partir de différents phénomènes physiques, comme l'absorption des rayons X pour la radiographie et le scanner, la résonance magnétique nucléaire pour l'IRM, les ultrasons pour l'échographie ou encore la radioactivité pour la médecine nucléaire [11].

1.3 Diagnostic médical

Le diagnostic médical est le processus par lequel un clinicien identifie la nature et la cause d'une pathologie chez un patient. Il constitue une étape essentielle de la prise en charge, car il oriente les décisions relatives au traitement, au suivi et à l'évolution thérapeutique [11]. Dans la pratique clinique, le diagnostic ne repose généralement pas sur une seule source d'information. Le médecin confronte plusieurs types de données, notamment les symptômes du patient, ses antécédents, les résultats biologiques, les images médicales et les comptes rendus spécialisés, afin d'aboutir à une décision raisonnée [7]. L'imagerie médicale permet de fournir une information visuelle sur l'état anatomique ou fonctionnel du patient [11]. Les rapports de radiologie, quant à eux, apportent une information textuelle et sémantique issue de l'interprétation du spécialiste [12]. Ainsi, l'image montre ce qui est observé, tandis que le rapport explique et interprète ces observations dans un contexte médical. Cependant, de nombreux systèmes d'intelligence artificielle appliqués au diagnostic restent encore limités à une seule modalité, en exploitant uniquement l'image ou uniquement le texte. Cette approche unimodale peut limiter la capacité du modèle à représenter toute la richesse des informations disponibles dans le dossier médical du patient [7]. Ainsi, l'exploitation conjointe des images médicales et des rapports textuels permet de se rapprocher davantage de la démarche réelle du

médecin. Cette complémentarité entre l'information visuelle et l'information textuelle constitue la base des systèmes intelligents multimodaux, capables d'intégrer différentes sources de données pour produire une classification ou une aide au diagnostic plus fiable [7][12].

1.4 Image numérique

L'image est un moyen de représentation visuelle permettant de transmettre une grande quantité d'informations. Elle joue un rôle essentiel dans la communication, l'observation et l'analyse des objets, des scènes ou des phénomènes. En informatique, une image peut être considérée comme une fonction bidimensionnelle dont les valeurs représentent une intensité lumineuse ou une information de couleur [13]. Une image numérique est une représentation discrète d'une image réelle ou synthétique. Elle est généralement organisée sous forme d'une matrice de pixels, où chaque pixel correspond à une valeur numérique décrivant une intensité ou une couleur [13] [14]. Cette représentation permet le stockage, l'affichage, le traitement et l'analyse automatique des images par des systèmes informatiques.

Les caractéristiques des principaux types d'images numériques, incluant le nombre de canaux, la profondeur de codage et l'espace des valeurs de pixels, sont synthétisées dans le Tableau 1.1 [13] [14].

TAB. 1.1. *Types d'images numériques*

Type d'image numérique	Description	Codage courant	Valeurs possibles
Image binaire	Chaque pixel possède uniquement deux états, généralement noir ou blanc.	1 bit par pixel	0 ou 1
Image en niveaux de gris	Chaque pixel représente une intensité lumineuse.	8 bits par pixel	0 à 255, soit 256 niveaux
Image couleur RVB	Chaque pixel est composé de trois composantes : rouge, vert et bleu.	3 canaux $\times$ 8 bits = 24 bits par pixel	$256^3 = 16\,777\,216$ couleurs

Au-delà de ces trois types fondamentaux, il existe d'autres représentations matricielles d'images. L'image indexée repose sur une matrice d'entiers (indices) associée à une palette de couleurs externe, ce qui optimise le stockage lorsque le nombre de couleurs est limité [14]. Enfin, sur le plan conceptuel, il est essentiel de distinguer les images matricielles (bitmap) des images vectorielles, qui sont décrites par des équations mathématiques et non par une grille de pixels [14].

Les caractéristiques importantes d'une image numérique peuvent être présentées comme suit :

- **Le pixel** : il constitue l'unité élémentaire d'une image numérique. Chaque pixel possède une position et une valeur d'intensité ou de couleur [13].

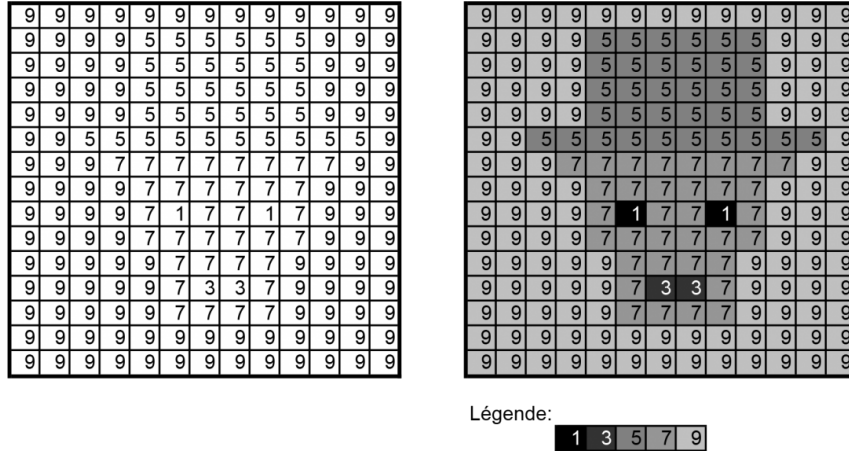


FIG. 1.1. Représentation d'une image numérique sous forme de matrice de pixels [15].

- **La dimension ou taille de l'image** : elle correspond au nombre de pixels selon les deux directions de l'image, généralement exprimé sous la forme largeur $\times$ hauteur. Par exemple, une image de 1920×1080 pixels contient 2 073 600 pixels [13].

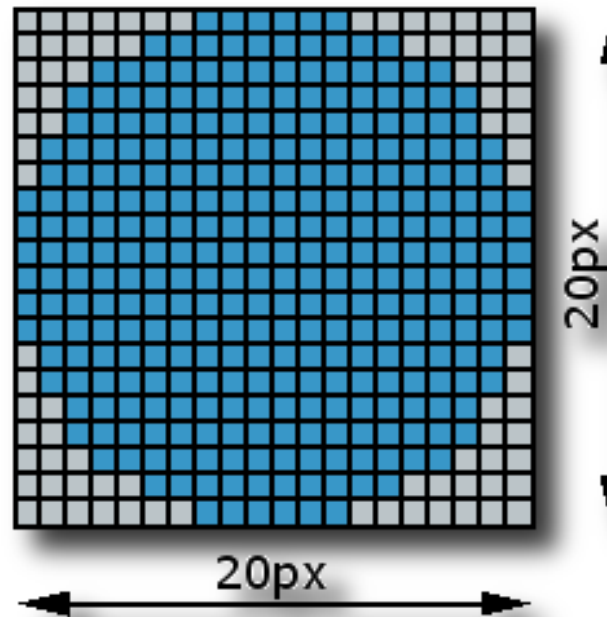


FIG. 1.2. Dimension d'une image numérique exprimée par sa largeur et sa hauteur en pixels [16].

- **La résolution spatiale** : elle indique la capacité d'une image à représenter les détails fins. Elle dépend de la densité des pixels ou des points par unité de distance et peut être exprimée, par exemple, en pixels par pouce (PPI), en DPI ou en pixels par millimètre [13].

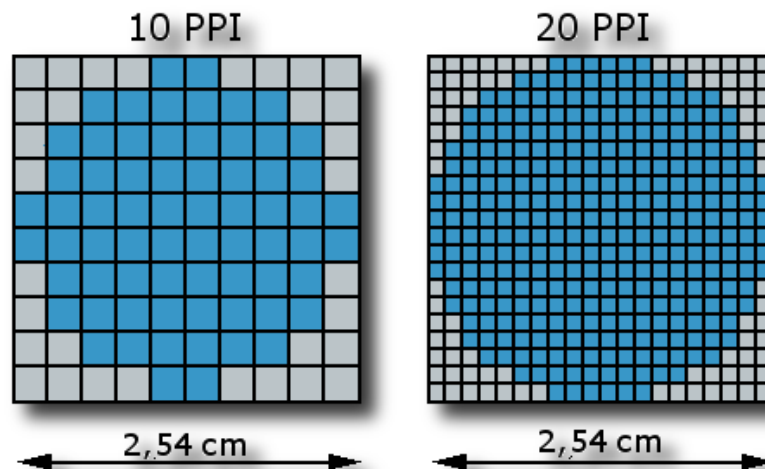


FIG. 1.3. Illustration de la résolution spatiale exprimée en pixels par pouce [16].

- **La luminance ou intensité** : elle exprime le niveau de luminosité d'un point ou d'une région de l'image. Dans une image numérique, elle correspond à la valeur d'intensité associée à un pixel ou à un ensemble de pixels [13] [14].

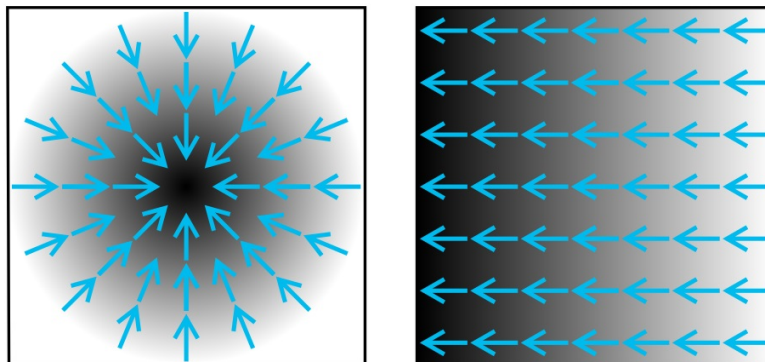


FIG. 1.4. Représentation des vecteurs gradients selon les variations de luminance : gradient radial (à gauche) et gradient linéaire (à droite) [17].

- **Le contraste** : il correspond à la différence ou à la variation de luminance/intensité entre deux zones voisines, ce qui permet de distinguer les objets, les détails ou les transitions dans l'image [13] [14]. Dans le cas simple de deux luminances extrêmes, il peut être mesuré par le contraste de Michelson [18] :

$$C = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}$$

où $L_{\max}$ et $L_{\min}$ représentent respectivement la luminance maximale et la luminance minimale.

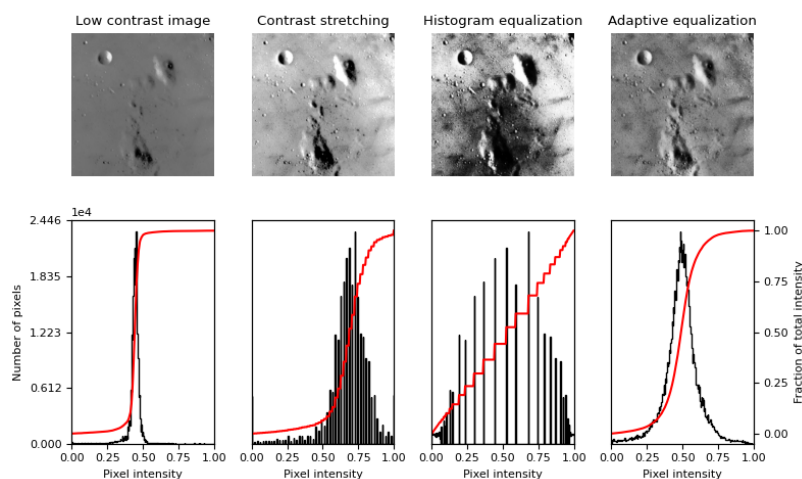


FIG. 1.5. Effet du contraste sur la visibilité des détails dans une image [19].

- **Les contours** : ils représentent les frontières entre les objets ou les régions d'une image. Ils apparaissent généralement lorsqu'il existe une variation importante de l'intensité ou de la couleur entre des pixels voisins [13] [14].

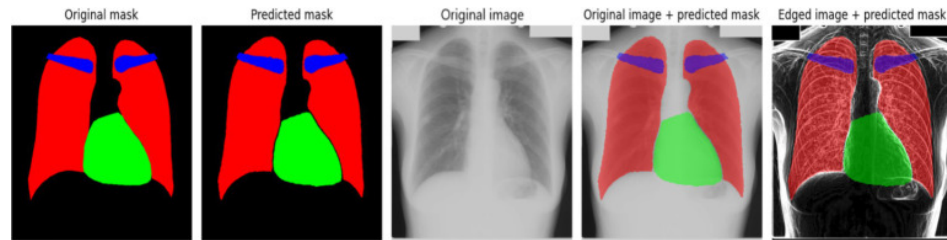


FIG. 1.6. Exemple de contours correspondant aux transitions d'intensité dans une image [20].

- **Les régions** : une région correspond à un ensemble connexe de pixels partageant une ou plusieurs propriétés communes, comme une intensité proche, une couleur similaire ou une texture homogène [13] [14].

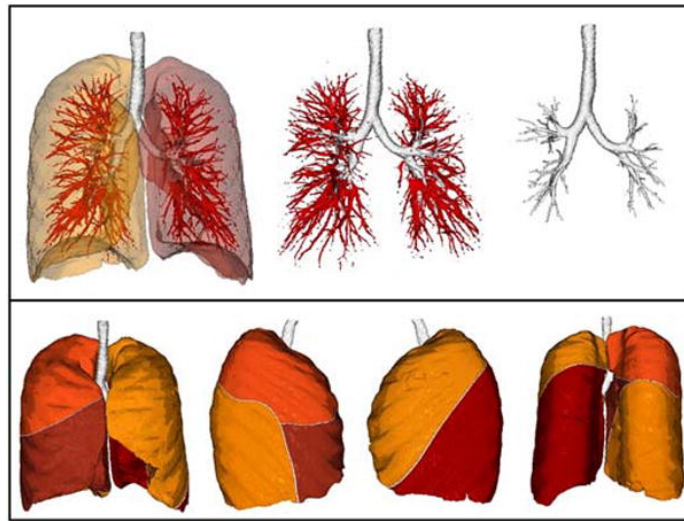


FIG. 1.7. Segmentation des régions et types pulmonaires : résultats du traitement automatique (en haut) et de l'outil interactif de segmentation lobaire (en bas) [21].

- **Le bruit** : il désigne des variations indésirables ou aléatoires des valeurs des pixels, pouvant dégrader la qualité de l'image [13] [14]. Comme illustré sur la **Figure 1.8**, la présence de bruit (Noisy) masque les détails fins de l'image originale. Pour y remédier, différents algorithmes de filtrage — tels que la variation totale (TV), le filtre bilatéral ou les ondelettes — sont généralement appliqués afin de restaurer l'information visuelle utile.

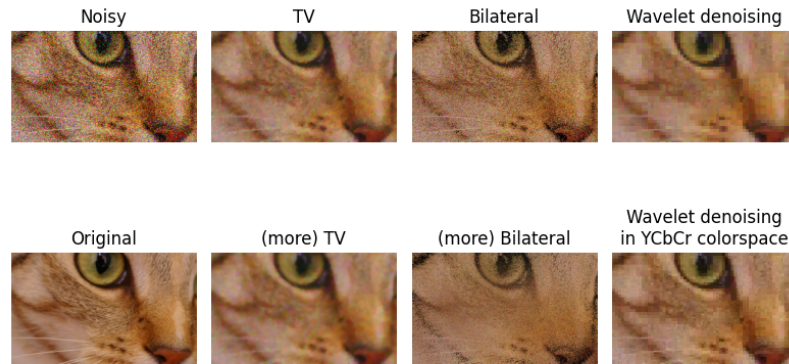


FIG. 1.8. Illustration de l'impact du bruit aléatoire sur une image et exemples de restauration [22].

- **La texture** : elle décrit l'organisation spatiale des intensités ou des couleurs dans une zone de l'image. Deux régions peuvent avoir des intensités globalement similaires mais présenter des arrangements spatiaux différents, ce qui produit des textures distinctes [13] [14] [23].

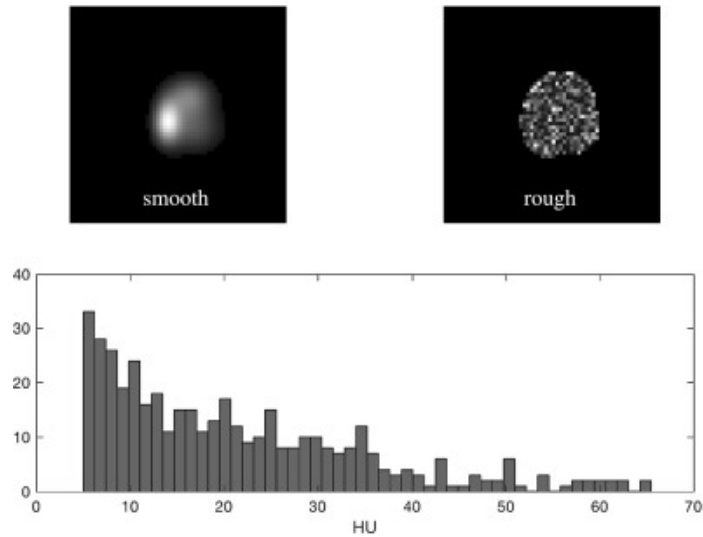


FIG. 1.9. Illustration de la notion de texture : deux régions peuvent posséder une distribution d'intensités similaire tout en présentant des organisations spatiales différentes [23].

- **L'histogramme** : il représente la distribution des valeurs d'intensité ou de couleur dans l'image [13] [14].

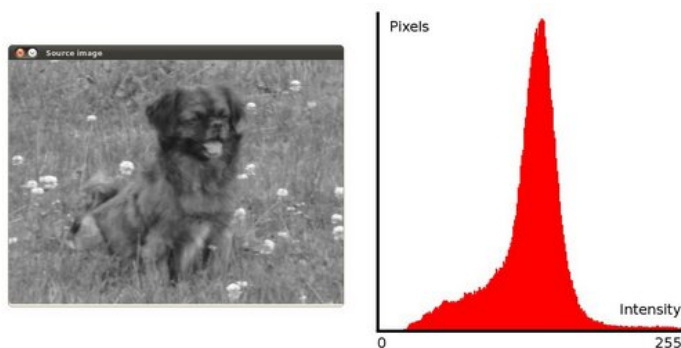


FIG. 1.10. Image en niveaux de gris et histogramme correspondant représentant la distribution des intensités des pixels [24].

La qualité d'une image numérique dépend de plusieurs facteurs, notamment la résolution spatiale, le contraste, le bruit, la netteté, le nombre de niveaux de gris ou de couleurs, ainsi que les conditions d'acquisition, de traitement et d'affichage [13] [14][25]. Dans le domaine médical, cette qualité est particulièrement importante, car elle influence l'interprétation visuelle des images et peut affecter la précision du diagnostic [25]. Elle peut également influencer la fiabilité et la reproductibilité des résultats obtenus par les systèmes d'analyse automatique, notamment lors de l'extraction de caractéristiques radiomiques [26].

1.4.1 Formats des images numériques

- **JPEG** : largement utilisé pour les photographies, repose sur une compression avec perte [27].
- **PNG** : permet une compression sans perte et prend en charge les images en niveaux de gris, en couleurs et la transparence [28].
- **GIF** : limité à 256 couleurs, peut être utilisé pour les images simples ou animées [29].
- **BMP** : format bitmap utilisé principalement dans l'environnement Windows, généralement peu compressé et souvent volumineux [30].

1.5 Imagerie médicale

L'imagerie médicale est née avec la découverte des rayons X à la fin du XIXe siècle et s'est progressivement développée grâce aux grandes avancées de la physique et de l'informatique. Aujourd'hui, elle regroupe un ensemble de techniques permettant d'acquérir, de traiter, de stocker et d'interpréter des images du corps humain à des fins diagnostiques, thérapeutiques ou de recherche biomédicale [31].



FIG. 1.11. Exemple de radiographie thoracique issue du dataset Indiana University Chest X-ray [32].

L'objectif principal de l'imagerie médicale est de produire une représentation visuelle compréhensible des structures internes du corps humain. Elle permet d'observer les os, les tissus, les organes, mais aussi certains aspects fonctionnels ou métaboliques. Elle joue ainsi un rôle essentiel dans le dépistage, le diagnostic, le suivi de l'évolution des maladies et l'évaluation de l'efficacité des traitements [11][31]. Les examens d'imagerie médicale peuvent fournir deux grands types d'informations [11] :

- **l'imagerie structurelle** : permet d'étudier l'anatomie des organes, leur taille, leur forme, leur localisation ou la présence d'une lésion.
- **l'imagerie fonctionnelle** : permet d'analyser le fonctionnement des organes, leur activité physiologique ou leur métabolisme.

Parmi les principales techniques d'imagerie médicale, on trouve :

- **Radiographie** : repose sur les rayons X (traversée différentielle selon la densité des tissus). Utilisée pour l'étude des structures osseuses, des articulations et de certaines pathologies pulmonaires [11][31].
- **Scanner (tomodensitométrie)** : utilise les rayons X pour obtenir des images en coupes fines, pouvant être reconstruites. Utilisé pour l'exploration des organes internes, la localisation d'anomalies et le suivi de pathologies [11][31].

- **Échographie** : technique basée sur les ondes ultrasonores (émission puis analyse des échos). Non invasive, sans rayonnements ionisants, permet une visualisation en temps réel [11][31].
- **IRM (imagerie par résonance magnétique)** : repose sur les propriétés magnétiques des noyaux d'hydrogène. Fournit des images très détaillées des tissus mous (cerveau, moelle épinière, muscles, organes internes) avec une excellente qualité de contraste [11][31].
- **Médecine nucléaire (scintigraphie, TEP)** : repose sur l'administration d'un traceur radioactif pour étudier le fonctionnement d'un organe ou l'activité métabolique d'une lésion. Utile en oncologie, cardiologie et neurologie [11][31].

Avec le développement de l'informatique médicale, les images issues des différents appareils doivent être stockées, échangées et traitées de manière standardisée. Dans ce contexte, la norme DICOM, pour Digital Imaging and Communications in Medicine, constitue le standard international utilisé pour la gestion, le stockage, la transmission et l'affichage des images médicales et des informations associées [33]. Aujourd'hui, l'imagerie médicale ne se limite plus à la simple acquisition d'images. Elle est de plus en plus associée aux méthodes d'intelligence artificielle, notamment pour l'aide au diagnostic, la détection automatique d'anomalies, la segmentation des organes et la classification des images médicales [2]. En radiologie, l'intégration d'outils d'IA constitue également un enjeu clinique important, car ces outils peuvent contribuer au diagnostic, à la quantification et à la gestion de certaines pathologies, à condition d'être évalués et surveillés correctement en pratique clinique [34].

1.6 Prétraitement et analyse des images médicales

Les images médicales obtenues à partir des différents dispositifs d'acquisition ne sont pas toujours directement exploitables par les systèmes informatiques. Elles peuvent contenir du bruit, présenter un faible contraste, avoir des tailles différentes ou provenir de conditions d'acquisition variées. Pour cette raison, une étape de prétraitement est souvent nécessaire avant toute analyse automatique [13, 14].

Le prétraitement regroupe l'ensemble des opérations visant à améliorer la qualité de l'image et à la rendre plus adaptée à l'interprétation visuelle ou au traitement automatique. Parmi les opérations les plus courantes, on trouve :

- le redimensionnement,
- la normalisation des intensités,

- la réduction du bruit,
- l'amélioration du contraste,
- et parfois la conversion de l'image en niveaux de gris.

Ces traitements permettent de réduire les variations inutiles entre les images et de faciliter l'extraction des informations pertinentes [13, 14].

L'analyse consiste ensuite à extraire des informations utiles à partir des images traitées. Elle peut inclure la détection de régions suspectes, la localisation d'anomalies, la segmentation d'organes ou de lésions, ainsi que la préparation des données pour des méthodes d'apprentissage automatique ou profond [13, 2]. Dans un système d'aide au diagnostic, cette analyse transforme l'image en une source d'information exploitable pour la prise de décision.

Contrairement aux approches classiques où le prétraitement et l'extraction manuelle de caractéristiques étaient centraux, les méthodes modernes d'intelligence artificielle délèguent cette extraction aux modèles eux-mêmes. Les réseaux de neurones convolutifs apprennent automatiquement des représentations visuelles adaptées à la tâche. Toutefois, cette capacité ne rend pas le prétraitement facultatif : la qualité des données d'entrée conditionne directement les performances du modèle.

Ainsi, le prétraitement et l'analyse constituent une étape clé pour préparer les données avant leur exploitation par des méthodes de classification. Ces deux notions assurent la transition entre les bases de l'imagerie médicale présentées dans ce chapitre et les approches d'intelligence artificielle qui seront abordées dans la suite de ce mémoire.

1.7 Compte rendu médical (Rapport)

Une autre catégorie importante de données médicales est constituée par les comptes rendus textuels, notamment les rapports de radiologie. Après un examen médical, le médecin spécialiste rédige un rapport en langage naturel décrivant les observations, les anomalies détectées, le diagnostic présumé et les recommandations pour la suite de la prise en charge [12]. Cependant, ces rapports sont souvent rédigés sous forme de texte libre, ce qui rend leur exploitation automatique plus difficile par les systèmes informatiques [12]. Le compte rendu d'un examen d'imagerie constitue une partie essentielle de l'acte d'imagerie médicale. Il contient généralement l'indication de l'examen, la technique utilisée, les résultats observés ainsi qu'une conclusion répondant à la question clinique posée. Il représente également une trace médico-légale et contribue à orienter la suite de la prise en charge du patient [35]. Ainsi, le rapport de radiologie ne se limite pas à une simple description de l'image ; il constitue une interprétation médicale destinée à faciliter la communication entre le spécialiste d'imagerie et

le médecin demandeur [35]. À titre d'illustration, voici un exemple de compte rendu extrait de la base de données Indiana Chest X-ray [36], que nous utilisons dans nos travaux :

« Findings : There is a focal airspace opacity in the right upper lobe. No pleural effusion is seen. Cardiac and mediastinal contours are normal. Impression : Right upper lobe pneumonia. »

Ce qui se traduit par :

« Résultats : Il existe une opacité focale de l'espace aérien dans le lobe supérieur droit. Aucun épanchement pleural n'est observé. Les contours cardiaques et médiastinaux sont normaux. Conclusion : Pneumonie du lobe supérieur droit. »

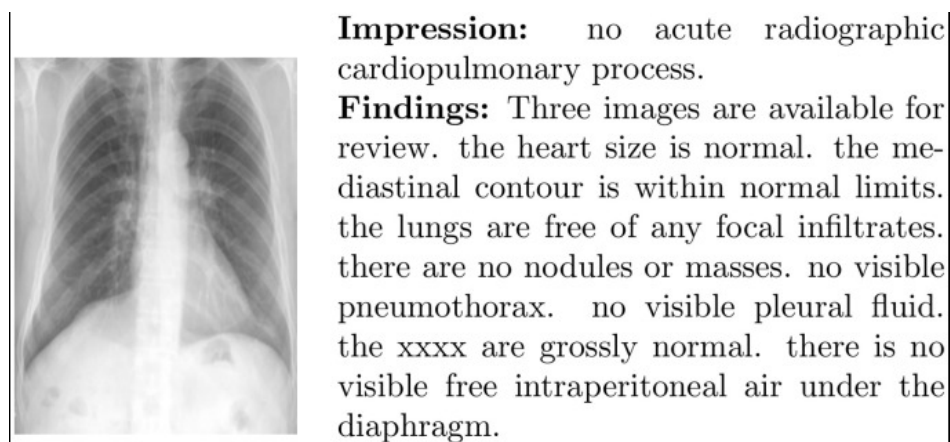


FIG. 1.12. Exemple de compte rendu radiologique réel : sections findings et impression d'une radiographie thoracique issue du dataset Indiana University Chest X-ray. [37].

Dans le cadre de nos travaux, nous exploitons ces comptes rendus comme seconde modalité, en complément des images radiologiques (vues frontale et latérale). L'objectif est de développer un modèle multimodal combinant un réseau de neurones convolutif (CNN) pour l'analyse des images et un transformeur (BERT) pour l'analyse textuelle, afin de classifier la pathologie ou l'anomalie détectée.

1.8 Conclusion

Dans ce chapitre, nous avons présenté les notions fondamentales liées aux images numériques, notamment leur définition, leurs types, leurs caractéristiques et leurs formats. Nous avons ensuite abordé l'imagerie médicale, qui permet de visualiser les structures internes du corps humain grâce à différentes techniques telles que la radiographie, le scanner, l'échographie, l'IRM et la médecine nucléaire.

Par ailleurs, nous avons mis en évidence l'importance du prétraitement et de l'analyse des images médicales, des étapes indispensables pour standardiser les données, réduire le bruit et préparer les entrées avant leur exploitation par des algorithmes de décision.

Enfin, nous avons enrichi cette étude par l'introduction d'une seconde modalité essentielle : les comptes rendus médicaux textuels, qui constituent une mine d'informations cliniques cruciales pour le diagnostic.

L'ensemble de ces notions représente une base essentielle pour la suite de ce mémoire. En effet, nos travaux s'intéressent à l'utilisation de l'apprentissage profond pour l'analyse et la classification multimodale des données médicales, combinant à la fois l'analyse visuelle des radiographies par des réseaux de neurones convolutifs (CNN) et l'extraction sémantique des rapports textuels par des modèles de type Transformeur (BERT).

Chapitre 2 : Classification multimodale des données médicales

2.1 Introduction

Les progrès récents de l'intelligence artificielle ont profondément transformé les méthodes d'analyse et d'interprétation des données médicales. Dans le domaine du diagnostic assisté par ordinateur, les systèmes intelligents sont aujourd'hui capables d'exploiter des volumes importants de données, de détecter des motifs complexes et d'aider les professionnels de santé dans la prise de décision [38] [1]. Cette évolution est particulièrement importante dans les situations où le diagnostic repose sur plusieurs sources d'information, comme les images médicales, les rapports cliniques, les antécédents du patient ou les résultats biologiques. Dans le cadre de ce mémoire, l'objectif principal est d'étudier une approche de diagnostic médical multimodal combinant deux types de données complémentaires : les images médicales et les rapports textuels associés. Les images médicales contiennent des informations visuelles essentielles sur l'état anatomique ou pathologique du patient, tandis que les rapports médicaux apportent un contexte clinique formulé en langage naturel. L'exploitation conjointe de ces deux modalités permet d'obtenir une représentation plus riche du cas étudié et s'inscrit dans l'évolution actuelle des systèmes d'aide au diagnostic multimodal [8][39].

Ce chapitre présente les concepts théoriques nécessaires à la compréhension de cette approche. Il commence par introduire l'intelligence artificielle en santé, puis le machine learning et le deep learning. Ensuite, il aborde la classification des données médicales, la classification d'images médicales par les réseaux de neurones convolutifs, puis l'extraction des caractéristiques visuelles par CNN. Le chapitre traite également du traitement automatique du langage médical, des Transformers et de BERT pour l'analyse des rapports médicaux. Enfin, il présente l'apprentissage multimodal, les stratégies de fusion image-texte et un état de l'art des approches basées sur CNN, BERT et les modèles multimodaux.

2.2 IA dans le domaine médical

L'intelligence artificielle, souvent désignée par l'abréviation IA, correspond à un ensemble de méthodes informatiques visant à concevoir des systèmes capables d'exécuter des tâches nécessitant habituellement une forme d'intelligence humaine. Ces tâches peuvent inclure la reconnaissance de formes, la compréhension du langage, la résolution de problèmes, l'apprentissage à partir de données ou encore la prise de décision [38]. Dans le domaine de la santé, l'IA occupe une place de plus en plus importante. Elle est utilisée pour l'aide au diagnostic, la prédiction de l'évolution des maladies, la segmentation d'organes ou de lésions, la classification d'images médicales, l'analyse des dossiers patients et l'extraction d'informations à partir de textes médicaux [1] [2]. Son intérêt principal réside dans sa capacité à analyser rapidement une grande quantité de données et à identifier des relations difficiles à percevoir

directement par l'observation humaine. L'IA ne remplace pas le médecin. Elle constitue plutôt un outil d'aide à la décision. Le diagnostic médical reste une tâche complexe qui nécessite l'expérience clinique, l'interprétation du contexte du patient et la validation par un professionnel de santé. Les systèmes d'IA peuvent cependant contribuer à réduire la charge de travail, attirer l'attention sur des anomalies, standardiser certaines analyses et améliorer la rapidité du processus diagnostique [1]. Les applications de l'IA en santé concernent plusieurs types de données. Les données visuelles, comme les radiographies, les scanners, les IRM ou les images de fond d'œil, sont souvent analysées par des modèles de vision par ordinateur. Les données textuelles, comme les comptes rendus médicaux ou les rapports radiologiques, sont traitées par des méthodes de traitement automatique du langage naturel. Les données tabulaires, comme les résultats biologiques ou les informations démographiques, peuvent également être intégrées dans des systèmes prédictifs [39] [40]. Dans un contexte médical réel, les décisions reposent rarement sur une seule source d'information. Un médecin interprète généralement une image en tenant compte du dossier clinique, des symptômes et du rapport associé. Cette réalité explique l'intérêt croissant des approches multimodales, qui cherchent à combiner plusieurs types de données afin d'améliorer la qualité du diagnostic [8][39].

2.3 Machine learning vs deep learning

Le machine learning est un sous-domaine de l'intelligence artificielle permettant à un système d'apprendre à partir de données sans être explicitement programmé [41]. Au lieu de définir manuellement les règles de décision, le modèle apprend à reconnaître des relations entre les entrées et les sorties à partir d'exemples. On distingue quatre types d'apprentissage : supervisé, non supervisé, semi-supervisé et par renforcement.

Le deep learning est une branche du machine learning fondée sur des réseaux de neurones profonds capables d'apprendre des représentations de plus en plus abstraites [42]. Contrairement au machine learning classique où l'extraction des caractéristiques est manuelle, le deep learning intègre cette étape directement dans le processus d'apprentissage [2], comme illustré par la Figure 2.1.

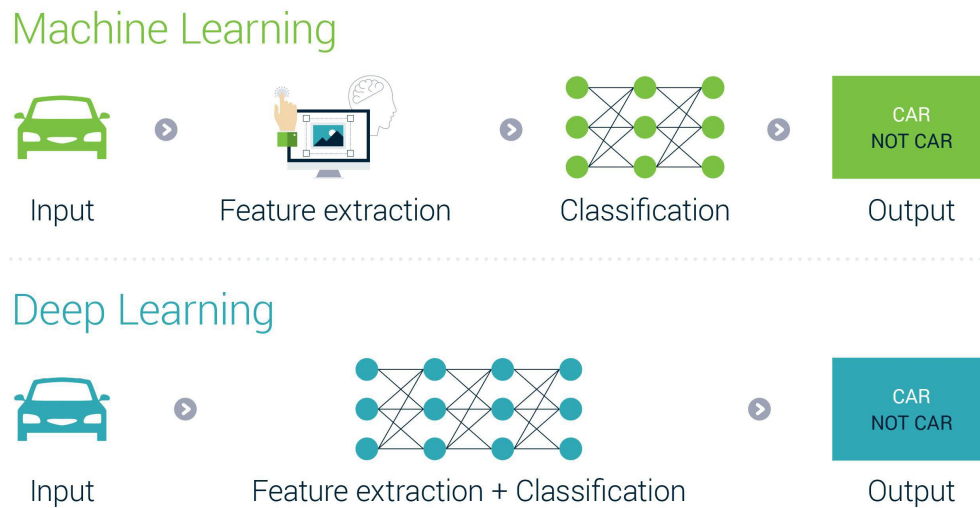


FIG. 2.1. Comparaison entre le machine learning et le deep learning : extraction manuelle versus automatique des caractéristiques [43].

En santé, le deep learning a montré de très bonnes performances dans plusieurs tâches, notamment la classification d'images médicales, la détection de lésions, la segmentation d'organes, l'analyse de textes médicaux et la fusion de données hétérogènes [1] [2] [40].

2.4 Classification des données médicales

La classification consiste à attribuer une ou plusieurs étiquettes à une donnée d'entrée. Trois types de classification sont distingués de manière générale en apprentissage automatique, chacun trouvant une application précise dans le contexte médical :

- La **classification binaire** associe chaque entrée à l'une de deux classes exclusives, par exemple la distinction entre un cas normal ou anormal.
- La **classification multi-classes** étend ce principe à plusieurs classes mutuellement exclusives, comme la distinction entre plusieurs stades de gravité d'une pathologie.
- La **classification multi-label** permet d'associer simultanément plusieurs étiquettes à une même entrée, ce qui correspond au cas clinique courant où un patient présente plusieurs anomalies à la fois — c'est le cadre retenu dans ce travail.

Comme l'illustre la figure 2.2, ces trois approches se distinguent par la structure de leurs sorties selon la complexité du problème à traiter.

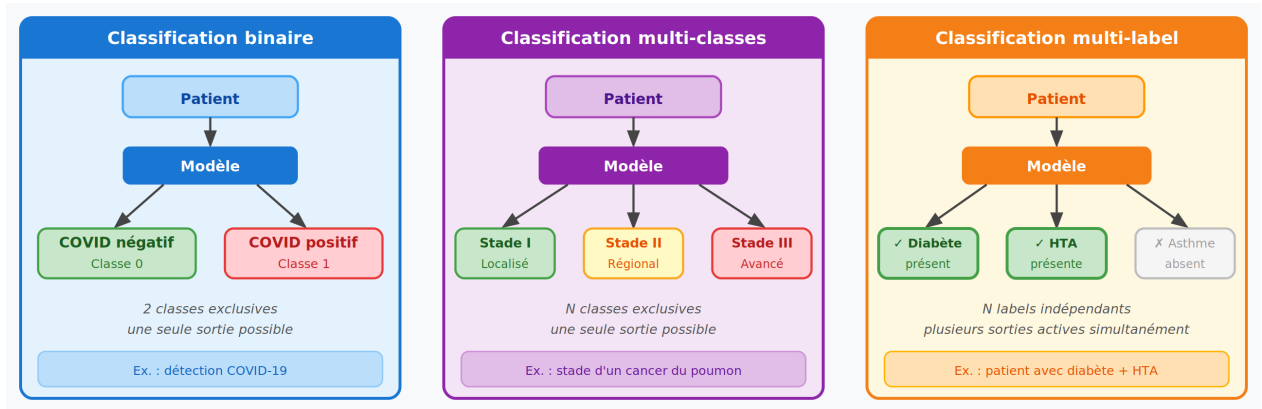


FIG. 2.2. Les trois types de classification en diagnostic médical.

La classification médicale soulève des défis spécifiques : déséquilibre des classes, données bruitées ou incomplètes, et coût élevé des erreurs.

Des métriques adaptées sont donc nécessaires au-delà de la simple exactitude : précision, rappel, score F1, sensibilité, spécificité et aire sous la courbe ROC. La sensibilité évalue la capacité à détecter les vrais positifs ; la spécificité évalue la capacité à rejeter les vrais négatifs. Pour obtenir une vision détaillée de ces erreurs de classification, l'utilisation de la matrice de confusion est indispensable, car elle cartographie précisément les correspondances entre les classes réelles et prédites. Enfin, à ces indicateurs s'ajoute le coefficient Kappa, une métrique statistique essentielle en imagerie médicale pour évaluer le niveau d'accord global entre les prédictions du réseau et la vérité terrain, en éliminant l'effet du hasard.

Dans une approche multimodale, la classification s'appuie sur une représentation fusionnée de plusieurs sources d'information, ce qui renforce la robustesse du modèle, chaque modalité apportant une information complémentaire [39][40].

2.4.1 Métriques d'évaluation des performances

L'évaluation d'un modèle de classification repose sur plusieurs métriques complémentaires permettant d'apprécier la qualité des prédictions. Dans le cas de la classification multilabel, ces indicateurs sont particulièrement importants, car ils fournissent une analyse plus représentative des performances que la seule exactitude globale. Les métriques de précision (*Precision*), de rappel (*Recall*) et de score F1 (*F1-score*), utilisées dans ce mémoire, sont calculées à partir de la matrice de confusion, composée des vrais positifs (TP), des faux positifs (FP), des faux négatifs (FN) et des vrais négatifs (TN).

- **Précision (Precision)** : elle mesure la proportion des prédictions positives qui sont effectivement correctes.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.1)$$

Une précision élevée indique que le modèle génère peu de faux positifs.

- **Rappel (Recall)** : également appelé sensibilité, il mesure la capacité du modèle à détecter les exemples réellement positifs.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.2)$$

Une valeur élevée traduit une bonne capacité de détection des cas positifs.

- **Score F1 (F1-score)** : il correspond à la moyenne harmonique entre la précision et le rappel.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.3)$$

Le score F1 fournit une mesure équilibrée des performances en tenant compte simultanément des faux positifs et des faux négatifs. Il est particulièrement adapté aux jeux de données déséquilibrés.

- **Courbe ROC et aire sous la courbe (AUC)** : la courbe ROC (*Receiver Operating Characteristic*) représente la sensibilité en fonction du taux de faux positifs pour différents seuils de décision. L'aire sous cette courbe (AUC) résume les performances globales du classifieur. Plus l'AUC est proche de 1, meilleure est la capacité du modèle à distinguer les classes positives des classes négatives.
- **Coefficient Kappa** : le coefficient de Cohen mesure le niveau d'accord entre les prédictions du modèle et les annotations de référence en corrigeant l'accord dû au hasard. Une valeur proche de 1 traduit un excellent accord, tandis qu'une valeur proche de 0 indique un accord comparable au hasard.

2.4.2 Classification d'images médicales par CNN

La classification d'images médicales vise à attribuer automatiquement une image à une catégorie diagnostique. Elle peut concerner différents types d'images, comme les radiographies, les images IRM, les scanners, les échographies, les images histologiques ou les images

rétiniennes. Cette tâche est essentielle dans les systèmes d'aide au diagnostic, car elle permet d'identifier rapidement des anomalies ou de proposer une prédiction préliminaire [2]. Les réseaux de neurones convolutifs, ou CNN, sont particulièrement adaptés au traitement des images. Contrairement aux réseaux entièrement connectés classiques, les CNN exploitent la structure spatiale des images. Ils utilisent des filtres de convolution capables de détecter automatiquement des motifs locaux tels que les contours, les textures, les formes ou les régions d'intérêt [44]. Un CNN est généralement composé de plusieurs types de couches. La couche de convolution applique des filtres sur l'image afin de produire des cartes de caractéristiques. La fonction d'activation, souvent ReLU, introduit de la non-linéarité dans le modèle. La couche de pooling réduit la taille spatiale des cartes de caractéristiques tout en conservant les informations importantes. Les couches entièrement connectées utilisent ensuite les caractéristiques extraites pour réaliser la classification finale. La dernière couche utilise généralement une fonction sigmoid pour une classification binaire ou softmax pour une classification multi-classes.

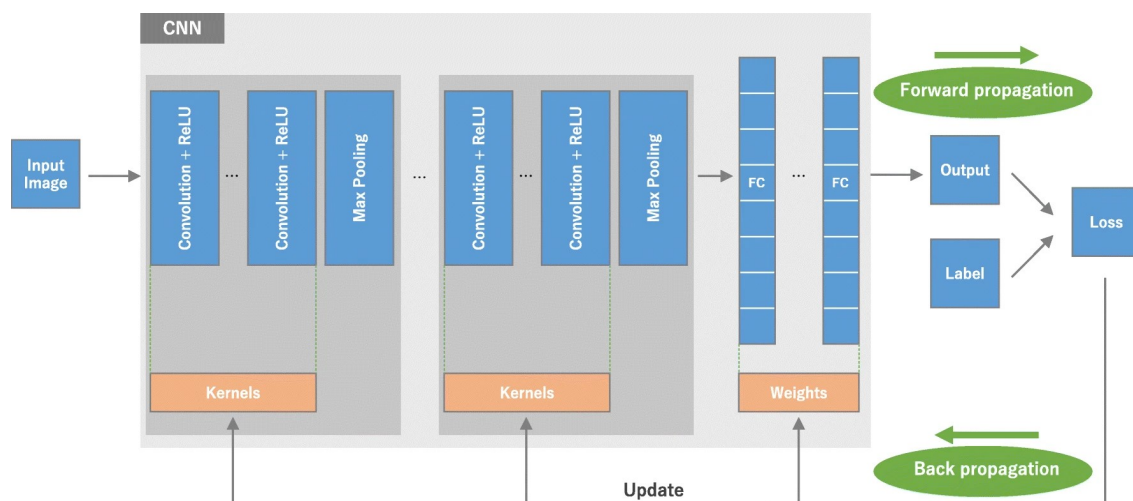


FIG. 2.3. Architecture générale d'un réseau de neurones convolutif [45]

Dans les images médicales, les CNN sont utilisés pour apprendre des caractéristiques visuelles associées à une pathologie. Par exemple, dans une radiographie thoracique, le modèle peut apprendre des motifs liés à une opacité pulmonaire. Dans une image de fond d'œil, il peut apprendre des signes liés à des lésions vasculaires. Dans une image histologique, il peut détecter des structures cellulaires anormales [2].

Plusieurs architectures CNN ont marqué l'évolution de la vision par ordinateur :

- **LeNet** : l'une des premières architectures CNN, utilisée pour la reconnaissance de chiffres manuscrits et de documents [44].

- **AlexNet** : a popularisé l'utilisation des CNN profonds après ses performances sur ImageNet [3].
- **VGG** : a montré l'intérêt d'utiliser des filtres de petite taille dans des réseaux plus profonds [46].
- **ResNet** : a introduit les connexions résiduelles permettant d'entraîner des réseaux très profonds [47].
- **DenseNet** : a proposé des connexions denses entre les couches pour favoriser la réutilisation des caractéristiques [48].
- **EfficientNet** : a cherché à améliorer le compromis entre précision et coût de calcul [49].

En médecine, l'apprentissage par transfert est souvent utilisé avec ces architectures. Il consiste à réutiliser un modèle pré-entraîné sur une grande base d'images, puis à l'adapter à une tâche médicale spécifique. Cette méthode est utile lorsque la base médicale disponible est limitée [2].

2.4.3 Extraction des caractéristiques visuelles avec CNN

Dans une approche basée sur CNN, l'extraction des caractéristiques visuelles est une étape centrale. Elle permet de transformer une image brute en une représentation numérique plus compacte et plus informative. Cette représentation, souvent appelée vecteur de caractéristiques, peut ensuite être utilisée pour la classification ou pour une fusion avec d'autres modalités. Les premières couches d'un CNN apprennent généralement des caractéristiques simples, comme les bords, les orientations et les contrastes. Les couches intermédiaires apprennent des motifs plus complexes, comme des textures, des régions homogènes ou des structures anatomiques. Les couches profondes apprennent des représentations plus abstraites, davantage liées à la tâche de classification [2][42]. Dans le cas d'un système multimodal image-texte, le CNN peut être utilisé comme extracteur de caractéristiques visuelles. L'image médicale est introduite dans le réseau, puis la sortie d'une couche profonde est récupérée sous forme de vecteur. Ce vecteur représente l'information visuelle pertinente de l'image. Il peut ensuite être concaténé ou combiné avec un vecteur textuel issu d'un modèle de langage fondé sur les Transformers, comme BERT [4][5], afin de réaliser une fusion multimodale [39]. L'extraction automatique des caractéristiques est un avantage majeur des CNN. Elle réduit la dépendance aux méthodes manuelles de prétraitement et permet au modèle d'apprendre directement les motifs utiles à partir des données. Toutefois, le prétraitement reste souvent important en

imagerie médicale. Selon le type d'image, il peut inclure le redimensionnement, la normalisation, l'amélioration du contraste, la suppression du bruit ou l'augmentation de données [2]. L'augmentation de données permet de créer artificiellement de nouveaux exemples à partir des images existantes. Elle peut inclure des rotations, des translations, des changements de contraste ou des retournements contrôlés. Cette technique aide à réduire le surapprentissage et à améliorer la généralisation du modèle.

2.4.4 Traitement automatique du langage (texte) médical

Le traitement automatique du langage naturel, ou NLP, regroupe les méthodes permettant aux systèmes informatiques d'analyser, comprendre et exploiter le langage humain [50].

Dans le domaine médical, il s'applique aux rapports cliniques, comptes rendus radiologiques, notes médicales, observations, diagnostics écrits et dossiers électroniques de santé [6] [51]. Le langage médical présente des particularités qui le rendent complexe à traiter. Il contient de nombreux termes spécialisés, des abréviations, des expressions propres à chaque spécialité, des formulations implicites et parfois des négations importantes. Par exemple, la phrase « absence de lésion visible » ne doit pas être interprétée comme la présence d'une lésion. La compréhension du contexte est donc essentielle.

Les méthodes classiques de NLP reposaient souvent sur des règles linguistiques, des dictionnaires médicaux ou des représentations simples comme le sac-de-mots (bag-of-words) ou TF-IDF. Ces approches peuvent être utiles, mais elles ne capturent pas toujours correctement le contexte et les relations entre les mots, comme l'illustre la figure 2.4. Avec l'apparition des modèles de deep learning, il est devenu possible d'apprendre des représentations plus riches du texte [4][5].

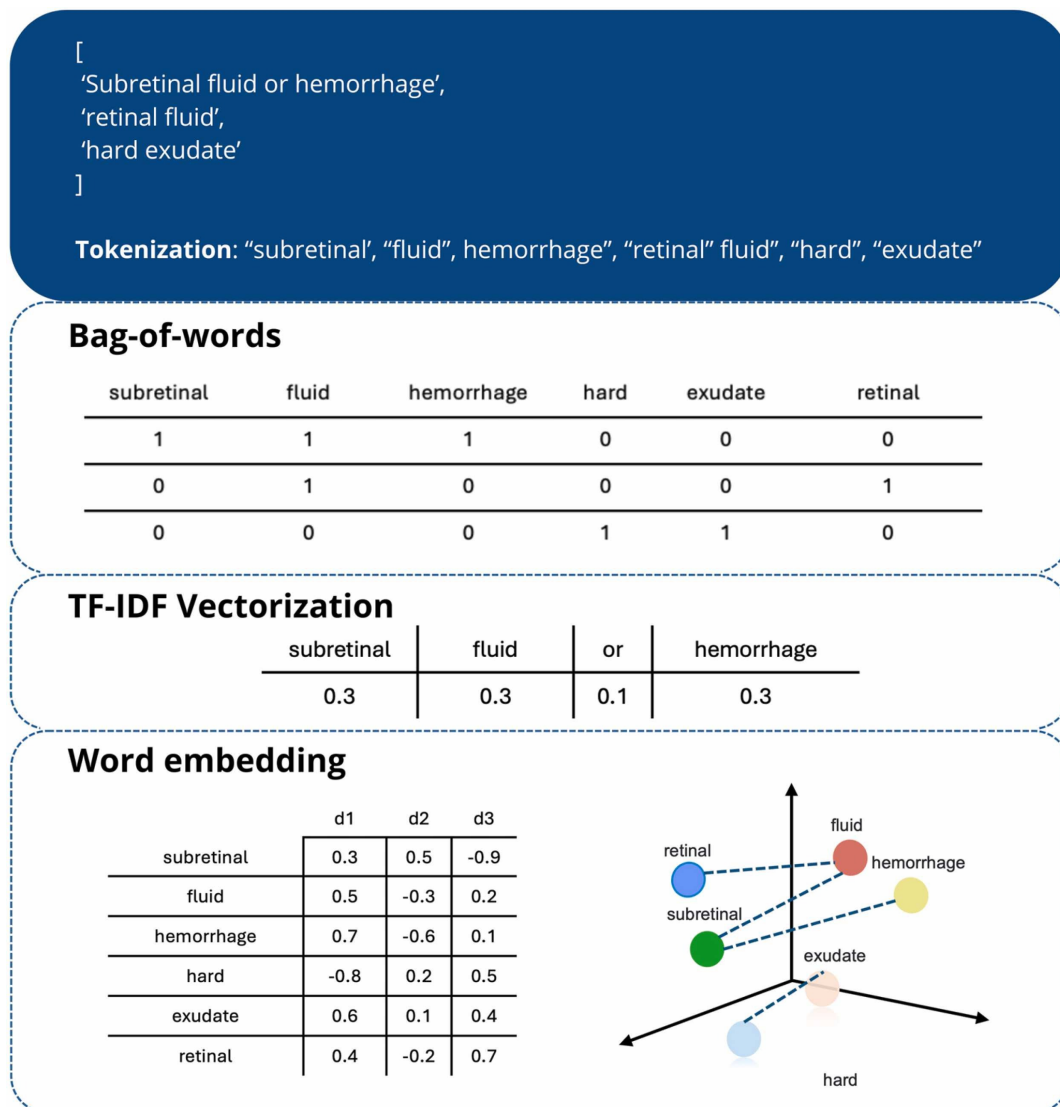


FIG. 2.4. Techniques de représentation textuelle en NLP médical : du sac-de-mots (Bag-of-Words) à TF-IDF et aux embeddings contextuels. Source : [50].

Dans un système de diagnostic multimodal, le rapport médical peut contenir des informations complémentaires à l'image. Il peut mentionner les symptômes, les observations du médecin, la localisation d'une anomalie, l'évolution de la maladie ou le contexte clinique. Ces informations textuelles peuvent aider le modèle à mieux interpréter l'image et à produire une classification plus fiable [39][40]. Le traitement automatique du langage médical peut donc être utilisé pour convertir un rapport écrit en vecteur numérique. Ce vecteur représente le contenu sémantique du texte et peut être combiné avec les caractéristiques extraites de l'image.

2.4.5 Transformers et BERT pour les rapports médicaux

Les Transformers sont des architectures de deep learning introduites pour le traitement des séquences, notamment le langage naturel [4]. Leur principe principal repose sur le mécanisme d'attention, qui permet au modèle d'évaluer l'importance relative des différents mots d'une phrase et de capturer des dépendances à longue distance, contrairement aux architectures récurrentes classiques. Comme l'illustre la Figure 2.5, l'architecture Transformer repose principalement sur un encodeur et un décodeur utilisant des mécanismes de Multi-Head Attention et des couches Feed Forward.

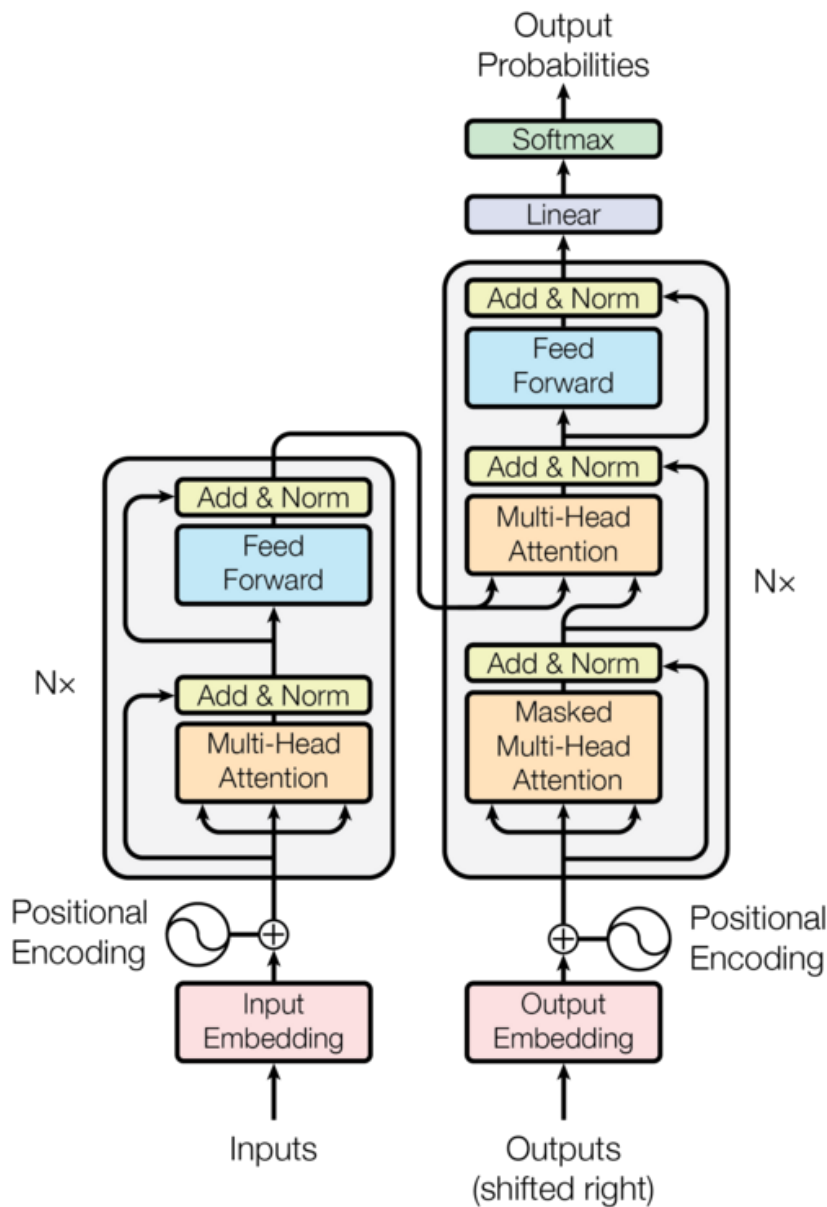


FIG. 2.5. Architecture du modèle Transformer [4]

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de langage fondé sur l'architecture Transformer [5]. Sa particularité est d'apprendre des représentations bidirectionnelles du texte, en tenant compte à la fois du contexte gauche et du contexte droit d'un mot, ce qui le rend très utile pour comprendre le sens d'un terme dans son contexte. Comme le montre la Figure 2.6, BERT est pré-entraîné à l'aide des tâches *Masked Language Modeling* (MLM) et *Next Sentence Prediction* (NSP), puis adapté à des tâches spécifiques par une étape de *fine-tuning* [5].

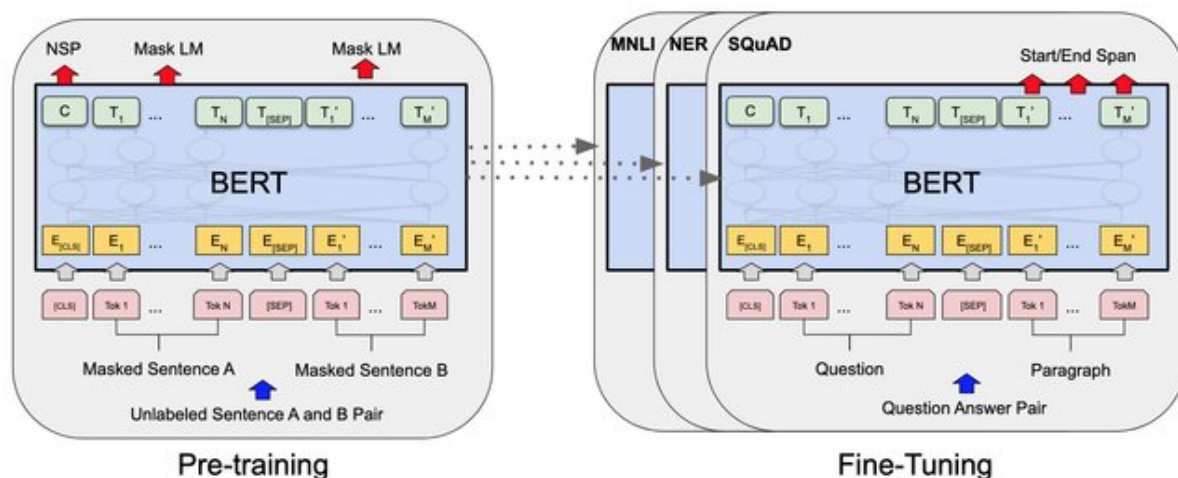


FIG. 2.6. Processus de pré-entraînement et de fine-tuning de BERT [5].

Dans le domaine médical, BERT peut être utilisé pour analyser des rapports cliniques ou radiologiques. Le texte est d'abord découpé en unités lexicales appelées *tokens*, puis transformé en vecteurs traités par les couches du modèle. La sortie de BERT peut être utilisée pour une classification textuelle ou comme représentation textuelle à fusionner avec une représentation visuelle [5]. Parmi les variantes spécialisées de BERT développées pour les textes biomédicaux et cliniques, on peut citer :

- BioBERT est pré-entraîné sur des corpus biomédicaux pour mieux représenter le vocabulaire scientifique et médical [51].
- ClinicalBERT est adapté aux notes cliniques et aux textes issus de dossiers hospitaliers [6].
- Bio_ClinicalBERT (initialisé avec BioBERT puis ré-entraîné sur les notes cliniques MIMIC-III) combine les forces des deux approches précédentes [52].
- GatorTron, plus récent, est entraîné sur plus de 90 milliards de mots issus de notes cliniques et d'articles biomédicaux, atteignant des performances supérieures pour l'extraction de concepts cliniques [53].

Ces modèles sont généralement plus pertinents que BERT généraliste lorsque les données contiennent un vocabulaire médical spécialisé. Pour un système basé sur CNN + BERT, le CNN extrait les informations visuelles à partir de l'image, tandis que BERT extrait les informations sémantiques à partir du rapport médical. Les deux représentations peuvent ensuite être combinées afin de réaliser une prédiction diagnostique.

2.5 Approche multimodal en diagnostic médical

L'apprentissage multimodal consiste à exploiter plusieurs modalités de données dans un même système d'apprentissage. Une modalité correspond à une source ou à une forme particulière d'information, comme l'image, le texte, le son, les données tabulaires ou les signaux physiologiques [8][39]. En médecine, les données sont naturellement multimodales, car le diagnostic repose souvent sur plusieurs éléments complémentaires [39].



FIG. 2.7. Principales sources de données multimodales dans le domaine biomédical [39].

Dans le cas étudié dans ce mémoire, les deux modalités principales sont l'image médicale et le texte médical. L'image fournit une information visuelle directe sur l'état du patient, tandis que le texte fournit une description clinique ou interprétative. Ces deux modalités ne portent pas exactement la même information. Elles peuvent se compléter, se confirmer ou parfois compenser les limites de l'autre. L'intérêt de l'apprentissage multimodal est de construire une représentation globale plus complète. Par exemple, une image peut montrer une anomalie visuelle difficile à interpréter sans contexte, tandis que le rapport peut préciser

la localisation, la gravité ou les observations cliniques. Inversement, le texte peut être incomplet, et l'image peut fournir des indices supplémentaires. Les modèles multimodaux sont particulièrement intéressants pour le diagnostic médical, car ils se rapprochent davantage du raisonnement clinique humain. Le médecin ne se limite pas à une seule donnée ; il combine l'examen visuel, les informations textuelles, les antécédents et les résultats complémentaires. L'objectif d'un modèle multimodal est donc de reproduire partiellement cette logique d'intégration. Cependant, l'apprentissage multimodal présente aussi des défis. Les données doivent être correctement alignées, c'est-à-dire que chaque image doit correspondre au bon rapport. Les modalités peuvent avoir des formats très différents. Les données peuvent être manquantes ou déséquilibrées. Enfin, le choix de la méthode de fusion influence fortement les performances du système [39] [40].

2.6 Les types de fusion

La fusion multimodale désigne la manière dont les informations provenant de plusieurs modalités sont combinées dans un modèle. Dans le cas d'une approche image + texte, l'objectif est de fusionner les informations visuelles extraites par CNN avec les informations textuelles extraites par BERT. Cinq grandes stratégies de fusion multimodale sont généralement distinguées [54] :

- **La concaténation précoce (Early Concatenation)** : concaténation des vecteurs de caractéristiques issus de chaque modalité avant projection dans un espace commun
- **L'unité multimodale à portes (Gated Multimodal Unit, GMU)** : utilisation de portes apprises pour pondérer dynamiquement chaque modalité.
- **La fusion bilinéaire (Bilinear Fusion)** : combinaison des modalités par un produit élémentaire (produit de Hadamard), créant un mécanisme de consensus naturel.
- **Le transformeur conjoint (Joint Transformer)** : projection des caractéristiques sous forme d'une séquence traitée par un bloc complet de transformeur avec auto-attention multi-têtes.
- **La fusion de bas rang (Low-Rank Fusion, LRF)** : décomposition des interactions bilinéaires en facteurs de faible rang, offrant un compromis entre expressivité et efficacité computationnelle.

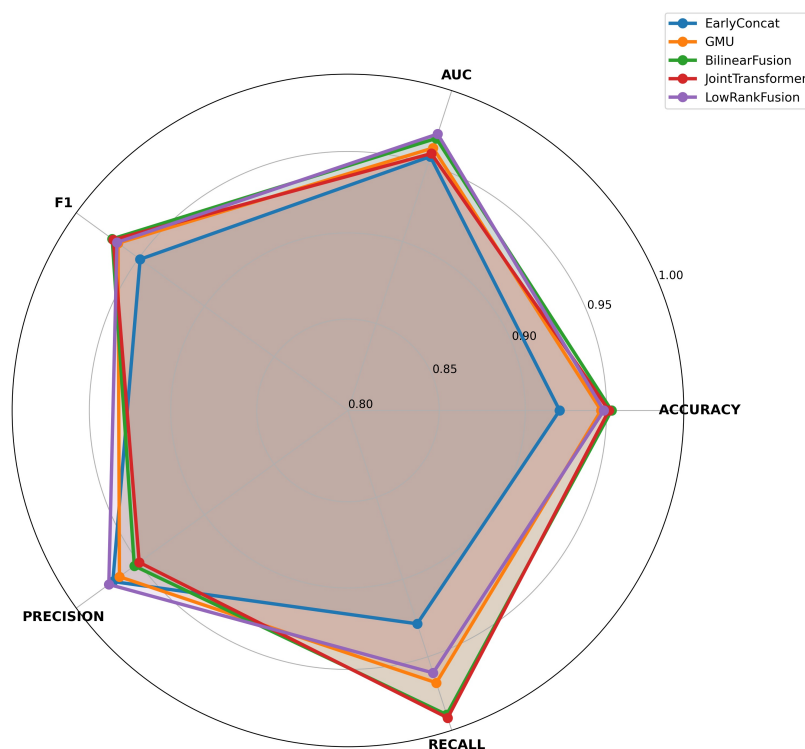


FIG. 2.8. Comparaison des performances des cinq stratégies de fusion multimodale sur l'ensemble de test [54]

2.7 État de l'art des approches multimodal (CNN, BERT)

Les approches basées sur les CNN ont largement dominé l'analyse automatique des images médicales au cours des dernières années. Elles ont été appliquées à la classification, la détection et la segmentation dans plusieurs domaines, notamment la radiologie, l'ophtalmologie, la dermatologie et l'histopathologie [2]. Les architectures comme VGG, ResNet, DenseNet et EfficientNet sont fréquemment utilisées comme bases de transfert learning pour extraire des caractéristiques visuelles à partir d'images médicales [46][47][48][49]. Dans les tâches de classification d'images médicales, les CNN permettent d'apprendre automatiquement les motifs visuels liés aux pathologies. Plusieurs travaux ont montré que ces modèles peuvent atteindre des performances élevées, parfois comparables à celles d'experts dans des conditions expérimentales contrôlées [1] [2]. Cependant, leur performance dépend fortement de la qualité des données, de la taille de la base, du prétraitement, de l'équilibre des classes et de la capacité du modèle à généraliser à des données provenant d'autres centres médicaux. Du côté textuel, les modèles Transformers et BERT ont amélioré les performances du traitement automatique du langage naturel. Le modèle BERT introduit par Devlin et al. a montré l'intérêt des représentations contextuelles bidirectionnelles [5]. Dans le domaine biomédical, BioBERT a

été proposé pour mieux traiter les textes scientifiques et médicaux [51]. ClinicalBERT a été développé pour représenter les notes cliniques et les textes hospitaliers [6].

Ces modèles ont été utilisés dans des tâches comme la classification de textes médicaux, l'extraction d'entités, la prédiction clinique et l'analyse de rapports. Les approches multimodales cherchent à dépasser les limites des modèles unimodaux. En diagnostic médical, une seule modalité peut être insuffisante. Une image peut contenir des signes visuels ambigus, tandis qu'un rapport peut contenir un contexte utile mais incomplet. La combinaison des deux permet d'améliorer la représentation du cas médical.

Plusieurs travaux récents ont concrètement mis en œuvre cette combinaison sur des données médicales réelles. Khader et al. ont proposé une architecture à base de transformeurs combinant radiographies thoraciques et paramètres cliniques tabulaires, démontrant un gain de performance par rapport aux modèles unimodaux équivalents [7]. Khan et Siddique ont mené une étude systématique de la fusion image-rapport sur le même corpus Indiana University Chest X-Ray que celui utilisé dans ce travail, combinant un ResNet-50 gelé à un DistilBERT gelé via une attention croisée légère ; leurs résultats montrent que le gain apporté par la fusion multimodale reste limité avec un encodeur textuel généraliste, et suggèrent qu'un encodeur biomédical spécialisé tel que BioBERT ou ClinicalBERT pourrait renforcer la synergie inter-modale [55]. Cette hypothèse est corroborée par Jahanian et al., dont le modèle RadFuse associe un Vision Transformer à ClinicalBERT via une attention croisée bi-directionnelle profonde sur les corpus MIMIC-CXR et IU X-ray ; leur étude d'ablation révèle que le retrait de l'encodeur clinique entraîne la dégradation la plus importante parmi tous les composants testés, confirmant le rôle déterminant d'un encodage textuel spécialisé dans la qualité de la fusion [56].

Au-delà de ces réalisations spécifiques, la littérature situe plus largement ce type de combinaison dans un cadre conceptuel de fusion multimodale. Duan et al. distinguent ainsi trois niveaux de fusion possibles : la fusion au niveau des données, qui combine les sources brutes avant tout traitement ; la fusion au niveau des caractéristiques, qui combine les représentations extraites séparément de chaque modalité – l'approche suivie par les travaux précédents – ; et la fusion au niveau des décisions, qui combine les prédictions produites indépendamment par chaque modalité [40].

Dans le domaine plus général de l'apprentissage multimodal, Baltrušaitis et al. ont formalisé les principaux défis de cette intégration (représentation, alignement, fusion proprement dite et co-apprentissage), un cadre qui structure encore aujourd'hui la majorité des architectures proposées en imagerie médicale et qui a depuis été repris par des revues centrées sur le diagnostic médical multimodal [8, 39].

Dans la pratique, ces architectures combinant image et texte suivent généralement un schéma commun : un réseau visuel extrait les caractéristiques de l'image, un modèle de

langage encode le rapport, puis les deux représentations sont fusionnées pour la classification. Certains travaux utilisent des CNN ou des Vision Transformers pour la partie image, et BERT, BioBERT, ClinicalBERT ou Sentence-BERT pour la partie texte. D'autres explorent des modèles vision-langage plus généraux, inspirés de CLIP, dont l'adaptation au domaine médical nécessite toutefois un apprentissage ou un ajustement sur des données spécialisées.

Au-delà du choix des encodeurs, Benomar et Gherbi [54] montrent que les performances d'un système multimodal dépendent également de la stratégie de fusion adoptée. Les cinq mécanismes de fusion qu'ils proposent, présentés en détail dans la section 2.6, ont été comparés expérimentalement afin d'évaluer leur efficacité selon différents compromis entre simplicité, expressivité et coût computationnel. Cette étude met en évidence qu'aucune stratégie n'est universellement optimale et que le choix du mécanisme de fusion dépend du contexte d'application et des caractéristiques des données.

Ainsi, l'état de l'art montre une progression claire depuis les modèles unimodaux vers les approches multimodales. Les CNN restent des outils puissants pour l'extraction de caractéristiques visuelles, tandis que BERT et ses variantes spécialisées sont bien adaptés à la compréhension des rapports médicaux. Leur combinaison représente une approche pertinente pour construire un système de diagnostic médical plus complet.

2.8 Conclusion

Dans ce chapitre, nous avons présenté les principaux concepts nécessaires à la compréhension d'un système de diagnostic médical multimodal basé sur CNN et BERT. Nous avons d'abord introduit le rôle de l'intelligence artificielle en santé, puis les notions de machine learning et de deep learning. Nous avons ensuite présenté la classification des données médicales et l'utilisation des CNN pour la classification d'images médicales.

Nous avons également montré que les CNN permettent d'extraire automatiquement des caractéristiques visuelles pertinentes à partir des images. En parallèle, le traitement automatique du langage médical permet d'exploiter les rapports cliniques et les comptes rendus associés aux images. Les Transformers et BERT offrent une solution performante pour représenter le contenu sémantique de ces textes. L'apprentissage multimodal permet de combiner ces deux sources d'information afin d'obtenir une représentation plus complète du patient.

Parmi les stratégies de fusion présentées (concaténation précoce, GMU, fusion bilinéaire, transformeur conjoint et fusion de bas rang), les approches de fusion intermédiaire (GMU, fusion bilinéaire, transformeur conjoint, LRF) sont particulièrement adaptées à une architecture CNN + BERT, car elles permettent de transformer séparément les caractéristiques visuelles et textuelles avant de les combiner pour la classification finale. La concaténation

précoce, bien que plus simple, constitue une baseline robuste mais suppose une indépendance des modalités qui peut limiter la performance.

Le chapitre suivant pourra s'appuyer sur ces bases théoriques pour présenter l'architecture proposée, les données utilisées, les étapes de prétraitement, la méthode de fusion et l'évaluation expérimentale du modèle.

Chapitre 3 : Expérimentations, résultats et discussion

3.1 Introduction

Ce chapitre présente l'ensemble des expérimentations menées pour évaluer notre approche multimodale de classification thoracique multi-label. Après avoir détaillé la préparation du jeu de données Indiana University et la justification du regroupement des 118 labels initiaux en 12 catégories sémantiques, nous décrivons le pipeline complet : extraction visuelle par DenseNet121, extraction textuelle par BioClinicalBERT, fusion par concaténation simple, puis classification multi-label. Les résultats sont présentés sous forme de courbes d'apprentissage, métriques globales, matrices de confusion et analyse par groupe. Enfin, nous discutons les performances obtenues, les comparons contextuellement à d'autres stratégies de fusion (Low-Rank, Bilinear) et justifions pourquoi la concaténation simple constitue un choix pertinent dans notre contexte de dataset modérément volumineux.

3.2 Description et préparation des données

3.2.1 Jeu de données Indiana University Chest X-ray

Le jeu de données Indiana University Chest X-ray [36], disponible publiquement via la plateforme OpenI [57] et accessible sur Kaggle [32], est composé de trois éléments : un ensemble d'images radiographiques et deux fichiers CSV complémentaires.

Le fichier `indiana_reports.csv` contient 3 851 comptes rendus textuels par patient, comportant notamment les champs *findings* (observations radiologiques), *impression* (conclusion clinique) et *problems* (liste des pathologies séparées par des points-virgules, couvrant 118 labels distincts), comme l'illustre la figure 3.1. Les deux champs textuels sont concaténés sous la forme « *Findings : ..., Impression : ...* » afin de constituer le rapport clinique utilisé en entrée du modèle.

Le fichier `indiana_projections.csv` recense 7 466 radiographies correspondant aux 3 851 patients, soit en moyenne deux images par patient, et assure le lien entre chaque image et son compte rendu correspondant, en précisant l'incidence de prise de vue : frontale (51 %) ou latérale (49 %), comme le montre la figure 3.2.

Après prétraitement détaillé par la suite, 3 046 cas sont retenus pour l'expérimentation, avec 12 classes de pathologies thoraciques.

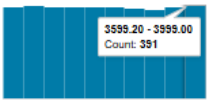
uid	Problems	image	findings	impression
	normal 36% No Indexing 2% Other (2380) 62%	Xray Chest PA an... 32% CHEST 2V FRONT... 2% Other (2541) 66%	[null] 13% The heart and lung... 1% Other (3286) 85%	No acute cardiopu... 8% No active disease. 3% Other (3423) 89%
1	normal	Xray Chest PA and Lateral	The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. T...	Normal chest x-XXXX.
2	Cardiomegaly;Pulmonary Artery	Chest, 2 views, frontal and lateral	Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lungs. Inferior...	No acute pulmonary findings.
3	normal	Xray Chest PA and Lateral		No displaced rib fractures, pneumothorax, or pleural effusion

FIG. 3.1. Extrait du fichier *indiana_reports.csv* : structure des comptes rendus textuels avec les champs *findings*, *impression* et *problems*.

uid	filename	projection
	7466 unique values	Frontal 51% Lateral 49%
1	1_IM-0001-4001.dcm.png	Frontal
1	1_IM-0001-3001.dcm.png	Lateral
2	2_IM-0652-1001.dcm.png	Frontal
2	2_IM-0652-2001.dcm.png	Lateral
3	3_IM-1384-1001.dcm.png	Frontal

FIG. 3.2. Extrait du fichier *indiana_projections.csv* : 7 466 radiographies réparties entre incidences frontales (51 %) et latérales (49 %).

3.2.2 Prétraitement des rapports textuels

Les comptes rendus sont nettoyés en concaténant les sections *findings* et *impression* après suppression des valeurs nulles. Les textes sont ensuite tokenisés avec le tokenizer de BioClinicalBERT, avec une longueur maximale fixée à 192 tokens (seuil couvrant 95 % des rapports).

Le Tableau 3.1 présente les étapes de nettoyage appliquées aux rapports textuels ainsi que leur impact sur la taille du jeu de données.

TAB. 3.1. Nettoyage des comptes rendus : concaténation des sections *findings* et *impression*

Étape	Description	Nombre
1	Dataset initial	3 851
2	Élimination des cas où <i>findings</i> ET <i>impression</i> sont nuls	-25
Dataset final		3 826

Comme indiqué le tableau 3.1, seuls les examens dont les sections *findings* et *impression* sont simultanément absentes ont été supprimés, ce qui réduit le nombre total de rapports de 3,851 à 3,826. Les données ainsi obtenues constituent la base sur laquelle seront appliquées les étapes de prétraitement décrites dans les sections suivantes.

3.2.3 Regroupement et nettoyage des labels

Le Tableau 3.2 résume les différentes étapes de nettoyage et de regroupement appliquées aux labels du jeu de données. Ces opérations permettent de réduire le nombre de classes de 118 à 12 catégories clinico-radiologiques plus pertinentes pour la tâche de classification multi-label.

TAB. 3.2. Nettoyage et regroupement des labels multi-label : réduction de 118 à 12 classes, et de 3 826 à 3 450 cas

Étape	Nb de classes	Nb de cas
Dataset initial	118	3 826
Exclusion des labels rares ou ambigus	-41	-376
Regroupement sémantique	-65	-
Dataset final	12	3 450

Comme l'indique le Tableau 3.2, le nombre de labels a été réduit de 118 à 12 classes clinico-radiologiques. Cette étape s'accompagne également d'une diminution du nombre d'examens disponibles, passant de 3,826 à 3,450 cas après l'exclusion des labels rares ou ambigus. Ce prétraitement permet d'améliorer la qualité des annotations et de limiter les problèmes liés au déséquilibre des classes lors de l'entraînement du modèle.

Justification du regroupement sémantique et clinique

Le regroupement s'appuie sur quatre critères :

1. compartiment anatomique
2. apparence radiographique
3. proximité clinique
4. proximité sémantique dans les rapports

Par exemple :

- Parenchymal Opacity / Airspace Disease rassemble opacity, infiltrate, consolidation, pneumonia.
- Cardiac Failure / Enlargement regroupe cardiomegaly, heart failure, pulmonary edema.
- Devices / Post-surgical regroupe cathéters, stents, implants (non pathologiques mais visibles).

Les labels exclus (ex. lung, heart, technical quality of image unsatisfactory, tuberculosis, sarcoidosis) ne sont pas jugés inutiles médicalement, mais ils ne constituent pas des cibles supervisées stables (trop rares, trop générales, hors champ thoracique ou trop ambiguës). Cette stratégie permet d'obtenir un jeu de 12 classes multi label équilibrées et interprétables.

TAB. 3.3. *Regroupement des classes pour la classification multimodale*

Groupe	Justification
Normal (1354)	Correspond aux examens sans anomalie majeure. Comparable au label No Finding utilisé dans CheXpert.
Parenchymal Opacity / Airspace Disease (634)	Regroupe les anomalies d'opacité du parenchyme pulmonaire : opacity, infiltrate, consolidation, pneumonia. Visuellement, ces labels produisent souvent des zones d'opacité pulmonaire ; textuellement, ils apparaissent dans des contextes proches.
Atelectatic Volume Loss (324)	L'atélectasie est liée à une perte de volume pulmonaire ; pulmonary atelectasis et volume loss décrivent donc le même mécanisme radiologique général.
Suite sur page suivante...	

Groupe	Justification
Pleural Abnormality (251)	Regroupe les anomalies de l'espace pleural : effusion, pneumothorax, hydropneumothorax, hemothorax, épaissement pleural, angle costo-phrénique. Ces signes concernent la plèvre et les espaces autour du poumon.
Cardiac Failure / Enlargement (370)	Cardiomegaly, heart failure, pulmonary edema, pulmonary congestion et pericardial effusion sont liés au cœur, à sa taille, ou aux conséquences pulmonaires de l'insuffisance cardiaque.
Aortic / Vascular Change (347)	Aorta, atherosclerosis, thoracic aorta, aneurysm : même compartiment vasculaire médiastinal visible sur RX.
Chronic Hyperinflation / COPD (180)	Emphysema, COPD, hyperlucent lung, bullous emphysema, bronchiectasis et cystic fibrosis décrivent des maladies chroniques avec hyperinflation, lucence ou altération bronchopulmonaire.
Granulomatous / Fibrotic Scar (760)	Calcified granuloma, granulomatous disease, cicatrix, fibrosis, interstitial lung disease : anomalies chroniques, cicatricielles ou calcifiées. Mayo Clinic note que les nodules pulmonaires calcifiés sont souvent liés à une ancienne infection résolue.
Mass / Nodule (119)	Mass et nodule sont des lésions focales pulmonaires. Ce regroupement est aussi proche du label Lung Lesion dans CheXpert.
Skeletal / Thoracic Deformity (439)	Fractures, scoliosis, kyphosis, spondylosis, deformity, pectus, osteoporosis : anomalies osseuses ou de la cage thoracique visibles sur CXR.
Devices / Post-surgical (371)	Cathéters, stents, tubes, sutures, implants, instruments : ce ne sont pas des maladies, mais des objets radiologiquement visibles et cliniquement utiles. CheXpert garde aussi une classe Support Devices.
Diaphragmatic / Hiatal (215)	Diaphragm, eventration, hiatal hernia, diaphragmatic hernia : même région anatomique et même type de signes basithoraciques/diaphragmatiques.

3.2.4 Fusion des données textuelles et images radiographiques

À partir du fichier `indiana_projections.csv`, nous filtrons les projections pour ne conserver que les vues *frontal* et *lateral*. Après suppression des doublons et pivotement de la table, nous fusionnons avec le jeu de données textuel (3 450 patients) et ne conservons que les patients disposant à la fois d'une vue frontale et d'une vue latérale.

Le tableau 3.4 résume l'impact de cette étape de filtrage sur le nombre de patients retenus.

TAB. 3.4. Réduction du nombre de patients après fusion avec les images radiographiques

Étape	Nombre de patients
Dataset après nettoyage textuel et regroupement des labels	3 450
Dataset final après fusion (vue frontale ET latérale)	3 046

Comme l’indique le tableau 3.4, cette opération réduit le dataset de 3 450 à 3 046 patients, soit une perte de 404 cas ne disposant pas simultanément des deux incidences radiographiques.

Le tableau 3.5 détaille l’évolution de la distribution des 12 classes avant et après cette fusion.

TAB. 3.5. Évolution de la distribution des classes avant et après fusion avec les images radiographiques

Groupe	Avant fusion (3 450 pat.)	Après fusion (3 046 pat.)	Évolution
Normal	1 354	1 232	−122
Granulomatous / Fibrotic Scar	760	666	−94
Parenchymal Opacity / Airspace Disease	634	535	−99
Skeletal / Thoracic Deformity	439	379	−60
Devices / Post-surgical	371	285	−86
Cardiac Failure / Enlargement	370	311	−59
Aortic / Vascular Change	347	293	−54
Atelectatic Volume Loss	324	282	−42
Pleural Abnormality	251	215	−36
Diaphragmatic / Hiatal	215	187	−28
Chronic Hyperinflation / COPD	180	160	−20
Mass / Nodule	119	107	−12

Comme le montre le tableau 3.5, la réduction est proportionnelle à la taille de chaque classe et ne modifie pas l’ordre de fréquence entre les groupes. La classe *Normal* enregistre la perte absolue la plus importante (−122 cas), tandis que *Mass / Nodule* reste la classe la moins représentée avec 107 cas après fusion. La distribution relative des classes est préservée, ce qui garantit que le déséquilibre initial n’est pas amplifié par cette étape de filtrage.

3.3 Méthodologie : pipeline multimodal proposé

Cette section présente l’architecture complète du modèle multimodal développé dans ce travail. Le pipeline intègre deux modalités complémentaires : les images radiographiques traitées par un encodeur visuel (DenseNet121) et les comptes rendus cliniques analysés par un encodeur textuel (BioClinicalBERT). Ces deux représentations sont ensuite fusionnées par concaténation simple avant d’être transmises à un classifieur fully connected produisant les

scores de classification pour les 12 groupes de pathologies thoraciques. La figure 3.3 illustre l'ensemble de ce pipeline.

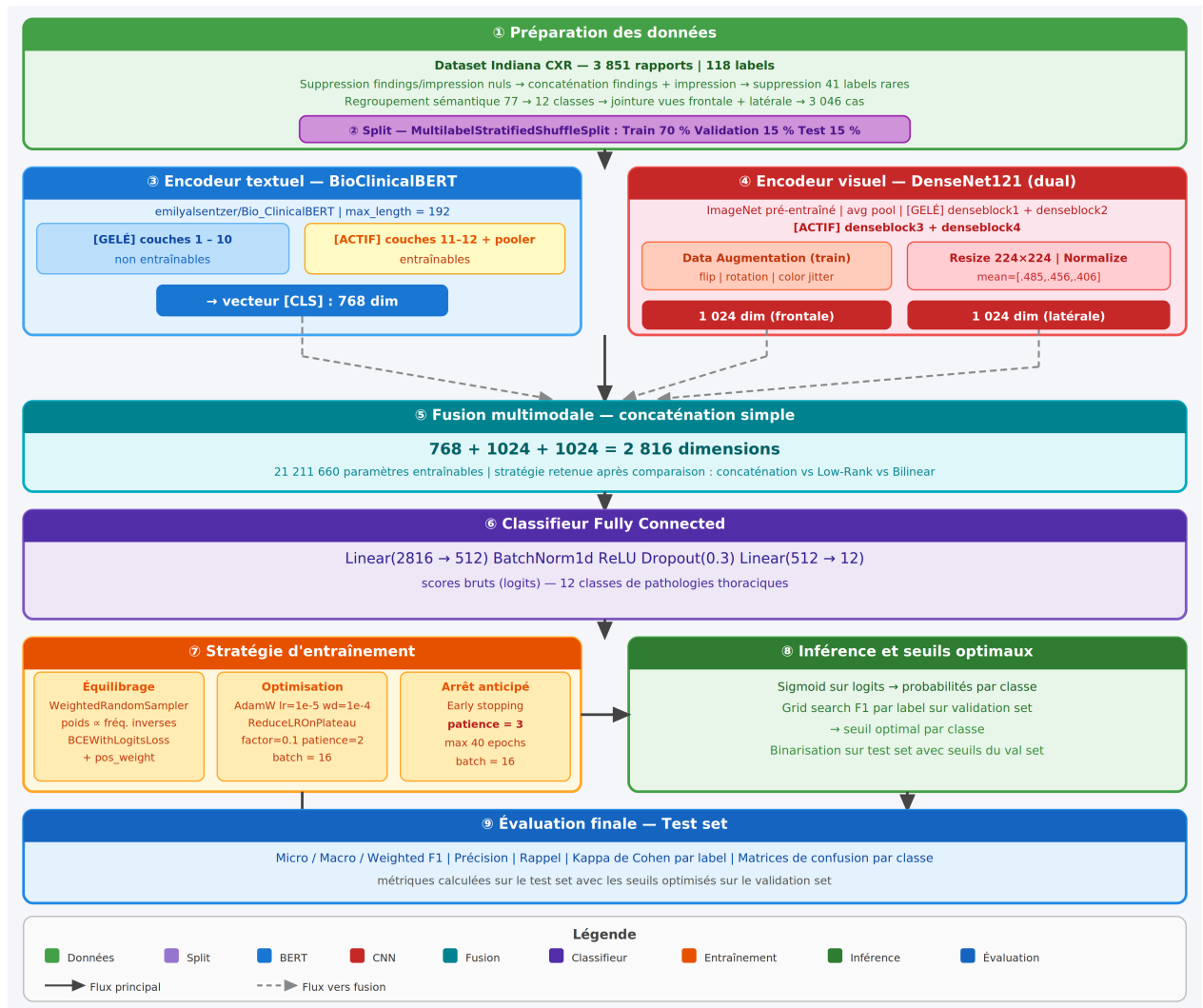


FIG. 3.3. Pipeline du modèle multimodal de diagnostic thoracique.

3.3.1 Explication de l'architecture proposée

L'architecture se compose de deux extracteurs de caractéristiques indépendants, suivis d'une fusion par concaténation et d'un classifieur fully connected. Les extracteurs sont initialisés avec des poids pré entraînés et partiellement figés pour éviter le surapprentissage.

A. Extraction des caractéristiques visuelles (DenseNet121) :

Pour la modalité visuelle, nous exploitons l'architecture DenseNet121 pré-entraînée sur le jeu de données ImageNet. Ce modèle est particulièrement adapté à l'imagerie médicale

grâce à ses connexions denses qui facilitent la propagation des caractéristiques primitives et maximisent le flux de gradients à travers le réseau.

Dans le cadre de notre pipeline multimodal final, la configuration optimale validée consiste à figer les deux premiers blocs d'extraction pour préserver les filtres génériques de bas niveau, et à rendre les deux derniers blocs denses (*denseblock3* et *denseblock4*) entièrement entraînaibles. Ce choix architectural majeur n'est pas posé *a priori* ; il découle d'une étude d'ablation systématique et d'une validation empirique par étapes détaillées à la section 3.5.3, au cours desquelles cette stratégie s'est avérée supérieure à toutes les autres configurations de gel testées.

La figure 3.4 illustre l'architecture générale du réseau DenseNet121 tel qu'utilisé dans notre pipeline.

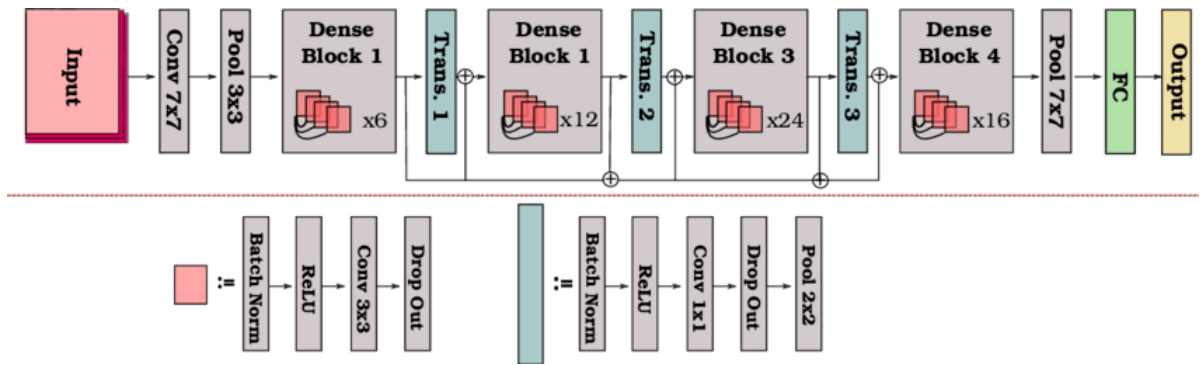


FIG. 3.4. Architecture générale du réseau DenseNet121 utilisée pour l'extraction des caractéristiques visuelles [58].

B. Extraction des caractéristiques textuelles (BioClinicalBERT) :

BioClinicalBERT est chargé avec des poids pré-entraînés sur corpus clinique. Seules les deux dernières couches de l'encodeur et le pooler sont affinés (les couches inférieures restent gelées). Le vecteur [CLS] final est utilisé comme représentation textuelle (768 dimensions).

Le choix de BioClinicalBERT comme encodeur textuel repose sur une adéquation directe entre le domaine de pré-entraînement du modèle et la nature des données exploitées dans ce travail. BioClinicalBERT est initialisé à partir de BioBERT [51], lui-même pré-entraîné sur des corpus biomédicaux issus de PubMed et PMC, puis ré-entraîné sur les notes cliniques de la base MIMIC-III [52]. Cette double spécialisation lui confère une représentation riche du vocabulaire radiologique, incluant les termes anatomiques, les abréviations cliniques et les formulations implicites caractéristiques des comptes rendus thoraciques.

Contrairement à un modèle BERT généraliste [5], BioClinicalBERT est capable de distinguer des nuances sémantiques importantes dans les rapports radiologiques, comme la différence entre *opacity* et *consolidation*, ou encore de traiter correctement les négations fréquentes dans les conclusions radiologiques (« *no pleural effusion* », « *heart size is normal* »). Ces distinctions sont cruciales pour une classification multi-label fiable.

La figure 3.5 présente l'architecture de BioClinicalBERT telle qu'utilisée dans notre pipeline.

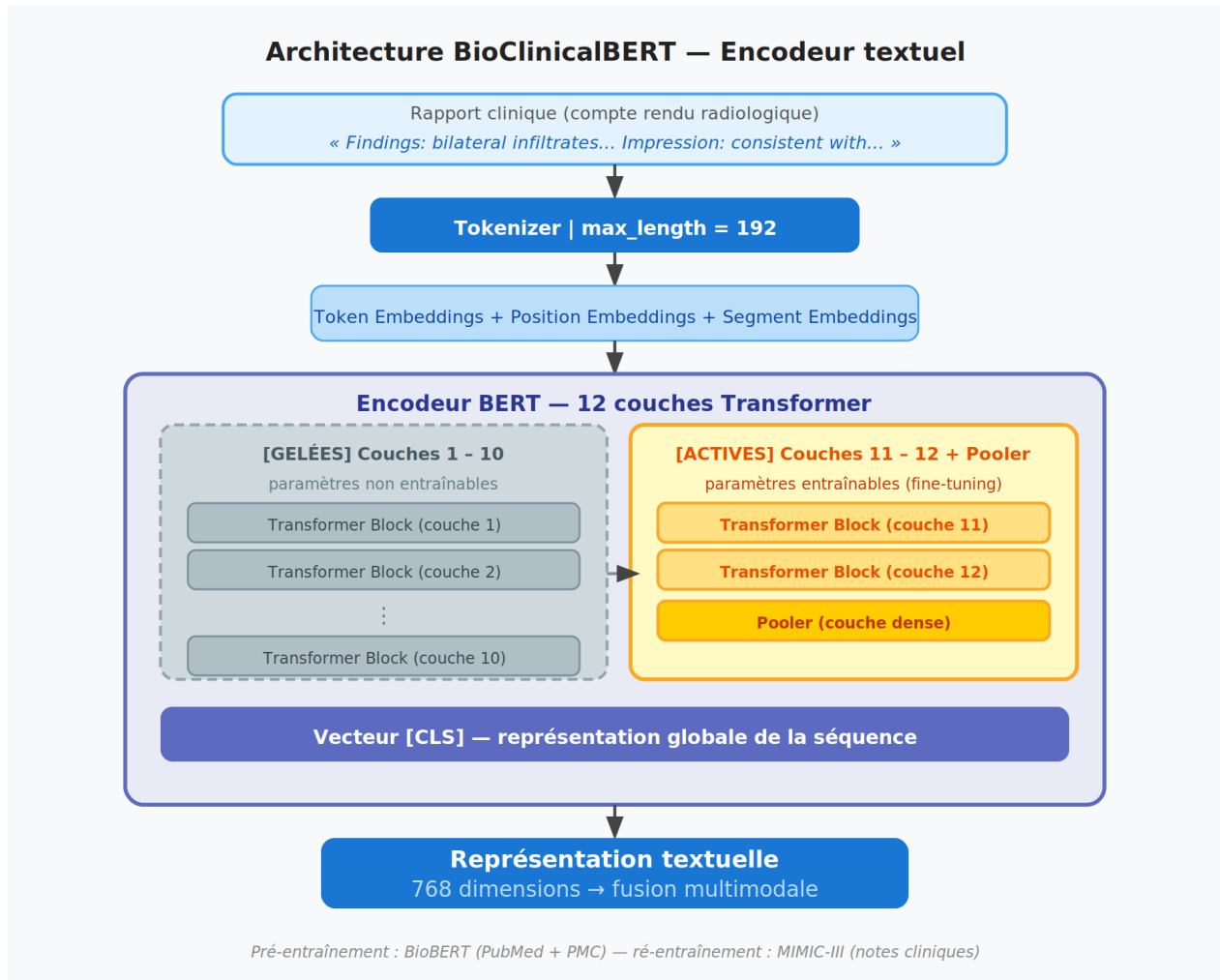


FIG. 3.5. Architecture de BioClinicalBERT utilisée comme encodeur textuel : couches 1-10 gelées, couches 11-12 et pooler affinés, vecteur [CLS] extrait comme représentation de 768 dimensions.

C. Fusion multimodale retenue : concaténation simple (Early Concatenation)

La fusion consiste à concaténer trois vecteurs de caractéristiques :

1. le vecteur textuel issu de BioClinicalBERT (768 dimensions),
2. le vecteur visuel de la vue frontale (1024 dimensions),
3. le vecteur visuel de la vue latérale (1024 dimensions).

Le vecteur fusionné obtenu est donc de dimension : $768 + 1024 + 1024 = 2816$ dimensions

Ce vecteur est ensuite passé dans un classifieur fully connected composé de :

- Une couche linéaire $2816 \rightarrow 512$
- BatchNorm
- ReLU
- Dropout(0.3)
- Une couche linéaire $512 \rightarrow 12$ (scores pour les 12 classes)

Justification du choix de la concaténation Bien que des fusions plus complexes (Low Rank, Bilinear, GMU, Transformer) aient montré leur supériorité sur de très larges volumes de données médicales, notre dataset est de taille modérée (environ 3046 paires image texte après filtrage). Dans ce contexte, la concaténation simple présente plusieurs avantages :

- Faible risque de surapprentissage
- Stabilité de l'entraînement
- Interprétabilité directe
- *Performances déjà très satisfaisantes (Micro F1 = 0,91)*

Les expériences préliminaires de comparaison (non systématiques) n'ont pas montré de gain significatif avec des fusions plus sophistiquées, ce qui conforte notre choix. L'utilisation de deux vues radiographiques (frontale + latérale) augmente naturellement la dimension de l'espace de fusion. La concaténation simple permet de préserver l'intégralité de l'information issue de chaque modalité et de chaque incidence, sans risquer de perte d'information ni de surapprentissage, ce qui est particulièrement adapté à un dataset de taille modérée.

D. Stratégie d'optimisation et de fine-tuning par blocs :

L'intégration conjointe des modalités visuelle (DenseNet121) et textuelle (BioClinical-BERT) au sein d'un même espace sémantique partagé impose de définir une politique de fine-tuning rigoureuse. Plutôt que d'appliquer un schéma de gel figé ou symétrique qui ignorerait les spécificités de chaque réseau, nous avons conçu une approche d'optimisation incrémentale et multimodale. La philosophie de cette démarche consiste à évaluer l'impact du dégel progressif et séquentiel des couches terminales de chaque modèle à travers une série de confrontations directes.

Cette stratégie d'apprentissage ciblé répond à trois impératifs majeurs :

1. **Spécificité sémantique** : Préserver le savoir généraliste des couches initiales des modèles fondateurs tout en libérant les blocs profonds (les couches d'attention terminales pour le texte et les blocs de convolution supérieurs pour la vision) responsables de l'encodage des concepts médicaux hautement spécialisés.

2. **Validation empirique par élimination** : Traiter chaque niveau de dégel comme une hypothèse scientifique mise à l'épreuve. Une configuration structurellement plus lourde n'est définitivement validée que si elle apporte un gain de performance statistiquement significatif par rapport à sa version restrictive.
3. **Équilibre paramétrique multimodal** : Ajuster la capacité d'apprentissage de l'extracteur visuel pour qu'il s'aligne de manière optimale avec la richesse sémantique issue de l'analyse textuelle, évitant ainsi qu'une modalité ne domine l'autre durant la fusion.

E. Équilibrage des classes et split des données :

Le jeu de données est divisé en 70% train, 15% validation, 15% test avec stratification multi label (itération de ShuffleSplit). Pour compenser le déséquilibre, un WeightedRandomSampler est utilisé dans le train, basé sur les poids inverses des fréquences de chaque label.

F. Hyperparamètres et stratégie d'entraînement :

- *Optimiseur* : AdamW ($lr = 1 \times 10^{-5}$, $weight\ decay = 1 \times 10^{-4}$)
- *Fonction de perte* : BCEWithLogitsLoss avec pondération des classes
- *Scheduler* ($factor = 0.1$, $patience = 2$) : — réduit le learning rate d'un facteur 0,1 si la loss de validation ne s'améliore pas pendant 2 époques consécutives.
- *Early stopping* : également activé avec une patience de 3 époques
- *Batch size* = 16
- Maximum 40 époques

3.4 Résultats expérimentaux

3.4.1 Courbes d'apprentissage (loss)

La figure 3.6 présente l'évolution de la perte (*BCEWithLogitsLoss*) sur les ensembles d'entraînement et de validation au cours des époques. Les deux courbes décroissent de manière régulière et convergente dès les premières époques : la perte d'entraînement part de 1,90 et la perte de validation de 1,20, toutes deux s'étalant autour de 0,25 après 35 époques. L'absence de divergence entre les deux courbes indique que le modèle ne souffre pas de surapprentissage. Le mécanisme d'arrêt anticipé (*early stopping*, patience = 3) a stoppé l'entraînement avant d'atteindre les 40 époques maximales autorisées, ce qui confirme que la convergence est atteinte naturellement. La validation loss reste légèrement inférieure à

la training loss sur la seconde moitié de l'entraînement, comportement attendu dans un contexte d'équilibrage par *WeightedRandomSampler* qui rend l'entraînement plus exigeant que l'évaluation.

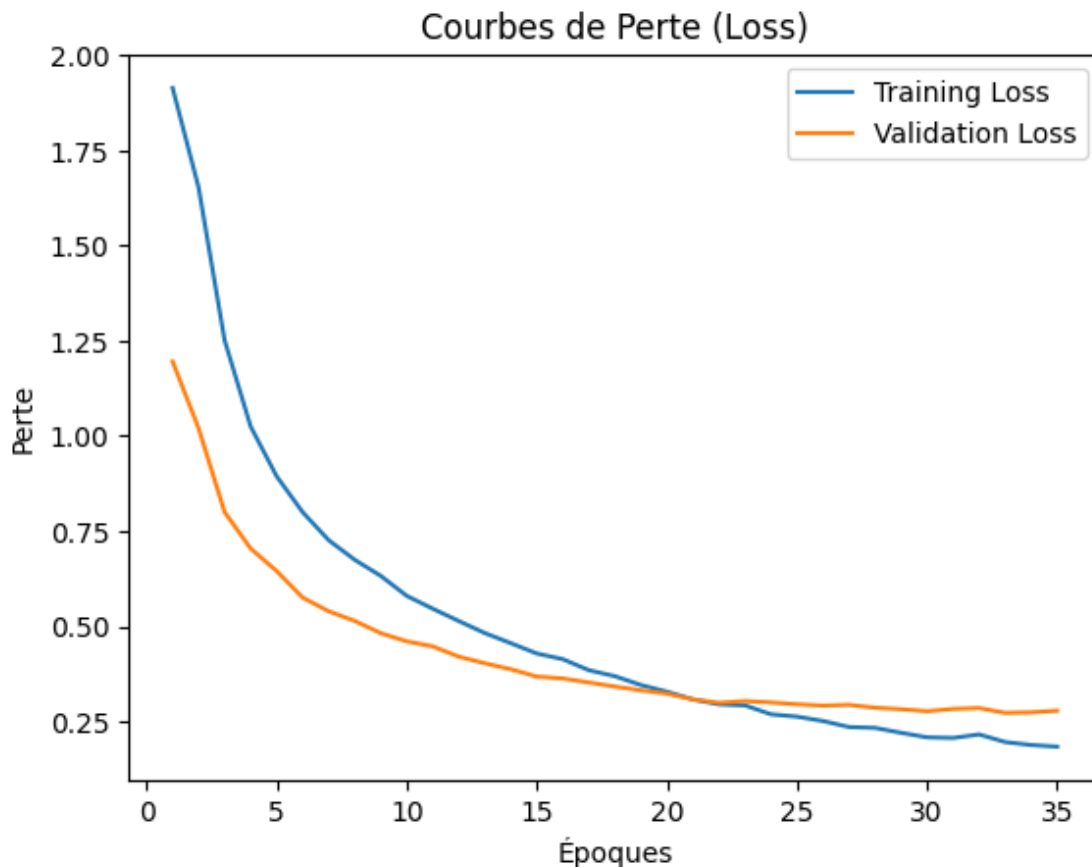


FIG. 3.6. Courbes d'apprentissage de la perte (*BCEWithLogitsLoss*) sur les ensembles d'entraînement et de validation

3.4.2 Performances globales

Le tableau 3.6 présente les métriques d'évaluation globales obtenues sur le test set. Le modèle atteint un Micro F1-score de 0,91, ce qui traduit une classification correcte de la grande majorité des paires image-texte. Le Macro F1-score de 0,90 confirme que les performances restent élevées et homogènes même sur les classes minoritaires telles que *Mass / Nodule* (support = 16) et *Chronic Hyperinflation / COPD* (support = 24). Le Weighted F1-score de 0,92 est cohérent avec le Micro F1, confirmant la robustesse du modèle face au déséquilibre des classes. Le Kappa moyen de 0,887 correspond à un accord presque parfait selon l'échelle de Landis & Koch, ce qui valide la qualité diagnostique des prédictions au-delà d'une simple correspondance statistique.

TAB. 3.6. Métriques d'évaluation globales du modèle multimodal sur le test set

Métrique	Micro	Macro	Weighted	Samples
Précision	0,91	0,90	0,92	0,93
Rappel	0,92	0,90	0,92	0,93
F1-score	0,91	0,90	0,92	0,92
Kappa moyen	0,887			

3.4.3 Matrices de confusion par label

La figure 3.7 présente les matrices de confusion individuelles pour chacune des 12 classes sur l'ensemble de test.

Les classes *Aortic / Vascular Change* et *Diaphragmatic / Hiatal* atteignent un rappel parfait ($R = 1,00$) avec 0 faux négatifs, traduisant une détection exhaustive de tous les cas positifs.

La classe *Normal* présente une excellente diagonale ($TN = 262$, $TP = 177$) avec très peu de confusions ($FP = 4$, $FN = 8$), reflétant la clarté visuelle et textuelle de cette catégorie.

Granulomatous / Fibrotic Scar affiche la précision la plus élevée de l'ensemble ($P = 0,989$, $FP = 1$), avec un seul faux positif sur 451 cas, ce qui confirme la distinctivité radiologique des lésions calcifiées et fibrotiques.

Parenchymal Opacity / Airspace Disease présente un rappel élevé ($R = 0,951$, $FN = 4$) avec une précision légèrement inférieure ($P = 0,885$, $FP = 10$), ce qui s'explique par la diversité des opacités pulmonaires regroupées dans cette classe.

Cardiac Failure / Enlargement et *Atelectatic Volume Loss* présentent un profil similaire : précision élevée mais rappel modéré (respectivement $R = 0,830$ et $R = 0,833$), avec des faux négatifs non négligeables ($FN = 8$ et $FN = 7$), reflétant la difficulté à détecter systématiquement ces pathologies dans des présentations atypiques.

Chronic Hyperinflation / COPD et *Mass / Nodule* sont les classes au support le plus faible (24 et 16 cas respectivement), mais le modèle maintient des précisions élevées (0,955 et 0,933), témoignant de l'absence de faux positifs excessifs malgré la rareté des exemples.

Pleural Abnormality présente un équilibre satisfaisant ($P = 0,848$, $R = 0,875$) avec peu de confusions ($FP = 5$, $FN = 4$), malgré la diversité des anomalies pleurales regroupées dans cette classe.

Devices / Post-surgical présente le nombre de faux positifs le plus élevé ($FP = 18$, $P = 0,69$), ce qui s'explique par la grande hétérogénéité visuelle de cette classe (cathéters, stents, implants, tubulures) dont les aspects radiologiques se confondent parfois avec des structures anatomiques normales.

Enfin, *Skeletal / Thoracic Deformity* cumule à la fois des faux positifs ($FP = 10$) et le nombre de faux négatifs le plus élevé ($FN = 12$), ce qui explique son F1-score le plus

faible (0,80) et reflète la variabilité morphologique des déformations osseuses ainsi que leur représentation textuelle moins standardisée dans les rapports cliniques.



FIG. 3.7. Matrices de confusion par label sur l'ensemble de test — modèle multimodal DenseNet121 + BioClinicalBERT (fusion par concaténation)

Le tableau 3.7 détaille les valeurs absolues de vrais positifs (TP), faux positifs (FP), faux négatifs (FN) et vrais négatifs (TN) pour chacune des 12 classes. Les classes *Aortic / Vascular Change* et *Diaphragmatic / Hiatal* présentent un FN nul, témoignant d'une détection exhaustive des cas positifs. À l'inverse, *Devices / Post-surgical* concentre le plus grand nombre de faux positifs (FP = 18), tandis que *Skeletal / Thoracic Deformity* cumule à la fois des faux positifs (FP = 10) et des faux négatifs (FN = 12), ce qui explique son F1-score plus

modeste (0,80). Les classes à faible support comme *Mass / Nodule* (TP = 14) et *Chronic Hyperinflation / COPD* (TP = 21) maintiennent néanmoins des scores de précision élevés (respectivement 0,933 et 0,955), ce qui atteste de la capacité du modèle à ne pas générer de faux positifs excessifs sur les classes rares.

TAB. 3.7. Métriques de performance par classe de pathologie (TP, FP, FN, TN)

Disease	TP	FP	FN	TN
Normal	177	4	8	262
Granulomatous / Fibrotic Scar	92	1	8	350
Parenchymal Opacity / Airspace Disease	77	10	4	360
Skeletal / Thoracic Deformity	45	10	12	384
Aortic / Vascular Change	44	3	0	404
Devices / Post-surgical	40	18	3	390
Cardiac Failure / Enlargement	39	1	8	403
Atelectatic Volume Loss	35	4	7	405
Pleural Abnormality	28	5	4	414
Diaphragmatic / Hiatal	28	2	0	421
Chronic Hyperinflation / COPD	21	1	3	426
Mass / Nodule	14	1	2	434

3.4.4 Courbes ROC et aires sous la courbe (AUC)

La figure 3.8 présente les valeurs d'AUC par classe et la figure 3.9 les courbes ROC correspondantes sur l'ensemble de test.

L'AUC moyen atteint 0,991, ce qui témoigne d'une excellente capacité discriminante du modèle sur l'ensemble des 12 pathologies, indépendamment du seuil de décision utilisé. Contrairement au F1-score qui dépend directement des seuils optimisés sur le validation set, l'AUC mesure la qualité intrinsèque de la séparation entre classes positives et négatives.

Onze classes sur douze dépassent un AUC de 0,985. Les classes *Aortic / Vascular Change* (AUC = 1,000) et *Diaphragmatic / Hiatal* (AUC = 0,999) atteignent une discrimination quasi parfaite, confirmant les observations faites sur le rappel parfait (R = 1,00) pour ces deux classes. *Mass / Nodule* obtient un AUC de 0,998 malgré un support très faible (16 cas), ce qui démontre la robustesse du modèle sur les classes rares. À l'opposé, *Skeletal / Thoracic Deformity* présente l'AUC le plus modeste (0,974), cohérent avec son F1-score le plus faible (0,80), en raison de la variabilité morphologique des déformations osseuses et de leur représentation textuelle moins standardisée dans les rapports cliniques.

--- AUC PAR CLASSE ---	
Aortic / Vascular Change	AUC = 0.9999
Diaphragmatic / Hiatal	AUC = 0.9990
Mass / Nodule	AUC = 0.9984
Granulomatous / Fibrotic Scar	AUC = 0.9975
Atelectatic Volume Loss	AUC = 0.9965
Pleural Abnormality	AUC = 0.9955
Normal	AUC = 0.9888
Devices / Post-surgical	AUC = 0.9881
Cardiac Failure / Enlargement	AUC = 0.9877
Chronic Hyperinflation / COPD	AUC = 0.9858
Parenchymal Opacity / Airspace Disease	AUC = 0.9852
Skeletal / Thoracic Deformity	AUC = 0.9741
AUC moyen : 0.9914	

FIG. 3.8. Valeurs d'AUC par classe sur l'ensemble de test — modèle multimodal DenseNet121 + BioClinicalBERT (AUC moyen = 0,991).

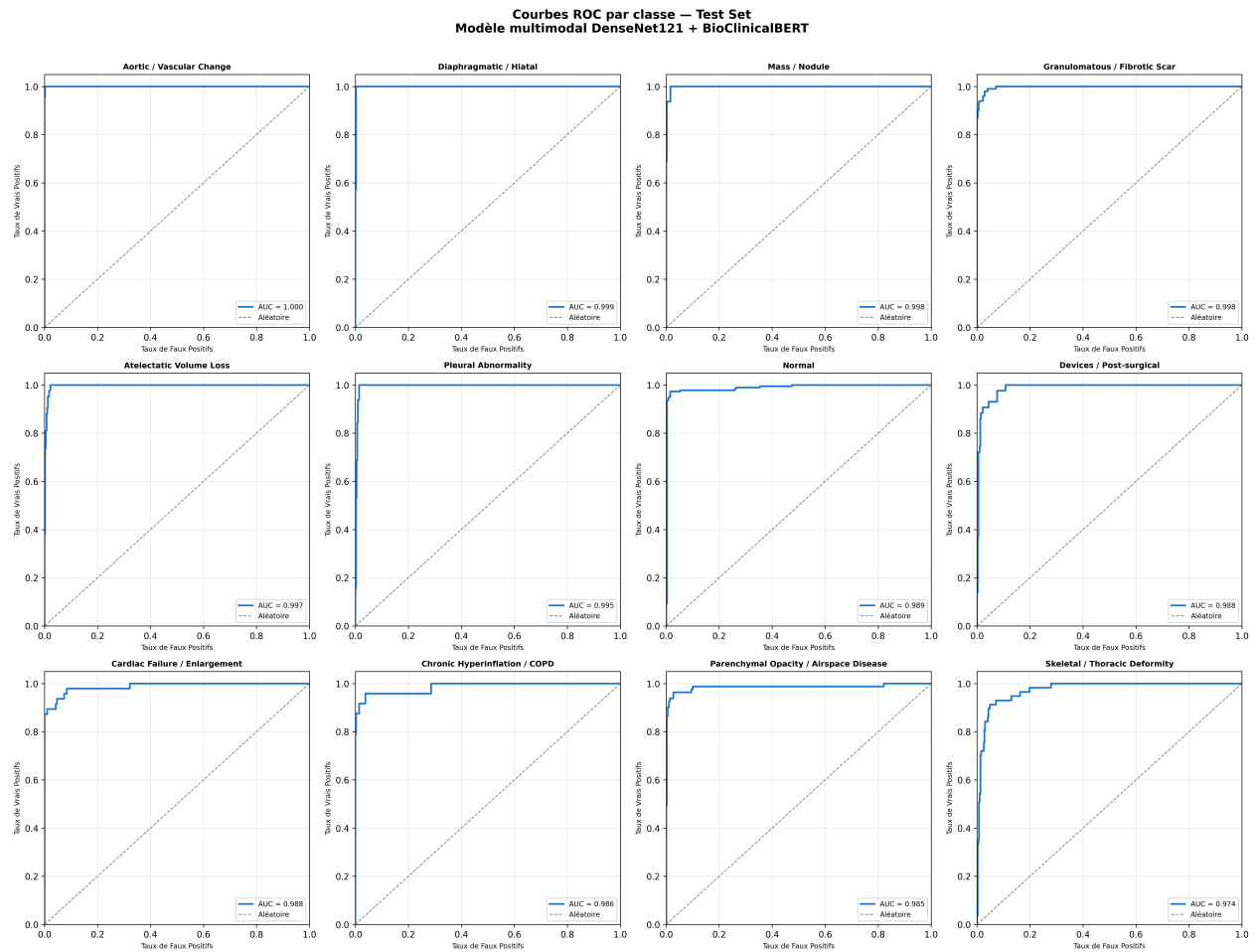


FIG. 3.9. Courbes ROC par classe sur l'ensemble de test — modèle multimodal DenseNet121 + BioClinicalBERT.

3.4.5 Analyse par groupe diagnostique

TAB. 3.8. Métriques détaillées par groupe de pathologies sur le test set

Groupe	Précision	Rappel	F1-score	Kappa
Aortic / Vascular Change	0,936	1,000	0,97	0,963
Diaphragmatic / Hiatal	0,933	1,000	0,97	0,963
Normal	0,978	0,957	0,97	0,945
Granulomatous / Fibrotic Scar	0,989	0,920	0,95	0,941
Chronic Hyperinflation / COPD	0,955	0,875	0,91	0,908
Mass / Nodule	0,933	0,875	0,90	0,900
Cardiac Failure / Enlargement	0,975	0,830	0,90	0,886
Parenchymal Opacity / Airspace Disease	0,885	0,951	0,92	0,898
Atelectatic Volume Loss	0,897	0,833	0,86	0,851
Pleural Abnormality	0,848	0,875	0,86	0,851
Devices / Post-surgical	0,690	0,930	0,79	0,767
Skeletal / Thoracic Deformity	0,818	0,789	0,80	0,776

L'analyse par groupe diagnostique révèle des disparités de performances qu'il convient d'interpréter à la lumière des caractéristiques propres à chaque classe.

Classes les mieux classées : Aortic/Vascular Change ($F1=0,97$, $\kappa=0,963$), Diaphragmatic/Hiatal ($F1=0,97$, $\kappa=0,963$) et Normal ($F1=0,97$, $\kappa=0,945$) affichent les meilleures performances. Ces classes bénéficient d'une représentation visuelle distinctive et d'une terminologie textuelle précise et cohérente dans les rapports radiologiques, ce qui facilite leur apprentissage par le modèle. Aortic/Vascular Change atteint notamment un rappel parfait ($R = 1,00$), signifiant que tous les cas positifs ont été correctement détectés.

Classes intermédiaires : Granulomatous/Fibrotic Scar ($F1=0,95$), Parenchymal Opacity/Airspace Disease ($F1=0,92$) et Chronic Hyperinflation/COPD ($F1=0,92$) présentent un bon équilibre entre précision et rappel, malgré une plus grande variabilité visuelle. La fusion multimodale joue ici un rôle important, car le texte clinique apporte un contexte sémantique complémentaire à l'image.

Classes les plus difficiles : Devices/Post-surgical ($F1=0,79$, $P = 0,690$) et Skeletal/Thoracic Deformity ($F1=0,80$, $R = 0,789$) obtiennent les scores les plus faibles. Pour Devices/Post-surgical, la faible précision ($P = 0,691$) indique un nombre élevé de faux positifs (18 FP), ce qui s'explique par la grande hétérogénéité visuelle de cette classe (cathéters, stents, implants). Pour Skeletal/Thoracic Deformity, le rappel modéré ($R = 0,789$) révèle une difficulté à détecter tous les cas positifs (12 FN), probablement due à la variabilité morphologique des déformations osseuses et à leur représentation textuelle moins standardisée dans les rapports.

3.5 Discussion

3.5.1 Interprétation des performances

Nos résultats montrent qu’une concaténation simple entre caractéristiques visuelles et textuelles permet d’atteindre des performances très satisfaisantes sur une tâche multi-label réaliste en imagerie thoracique. Le tableau ci-dessous résume l’ensemble des métriques obtenues sur le test set :

- **Micro F1 = 0.91** et **Samples F1 = 0.92** : le modèle classe correctement la grande majorité des paires image-texte, y compris les classes fréquentes.
- **Macro F1 = 0.90** : les performances restent élevées et homogènes même sur les classes minoritaires (Mass/Nodule, Chronic Hyperinflation/COPD).
- **Weighted F1 = 0.92** : cohérent avec le Micro F1, confirmant la robustesse du modèle face au déséquilibre des classes.
- **Kappa moyen = 0.887** : accord presque parfait entre les prédictions et les labels de référence selon l’échelle de Landis & Koch.

Sur le plan de la **précision et du rappel**, le modèle présente un bon équilibre global (précision micro = 0.91, rappel micro = 0.92). Par classe, on observe deux comportements distincts :

- Les classes à **haute précision et haut rappel**, comme Aortic/Vascular Change ($P = 0.936$, $R = 1.00$), Diaphragmatic/Hiatal ($P = 0.933$, $R = 1.00$) et Normal ($P = 0.978$, $R = 0.957$), bénéficient d’une représentation visuelle et textuelle claire et cohérente dans le dataset.
- Les classes plus **difficiles**, comme Devices/Post-surgical ($P = 0.690$, $R = 0.930$) et Skeletal/Thoracic Deformity ($P = 0.818$, $R = 0.789$), présentent soit un rappel plus faible (difficulté à détecter tous les cas positifs) soit une précision plus faible (faux positifs), ce qui s’explique par leur plus grande variabilité sémantique et visuelle.

Ces résultats confirment que l’approche multimodale DenseNet121 + BioClinicalBERT, avec une fusion par concaténation simple, constitue une base solide et compétitive pour la classification multi-label des pathologies thoraciques.

3.5.2 Comparaison contextuelle avec d’autres stratégies de fusion

Afin de valider le choix de la concaténation précoce comme mécanisme d’intégration sémantique, deux stratégies de fusion alternatives et plus sophistiquées ont été évaluées sous des conditions expérimentales rigoureusement identiques (BioClinicalBERT dégelé sur 2 couches et DenseNet121 dégelé sur le *denseblock4*) : la fusion de bas rang (*Low-Rank Fusion*) et la fusion bilinéaire (*Bilinear Fusion*). Les métriques de performance globales obtenues sur l’ensemble de test sont synthétisées dans le Tableau 3.9

TAB. 3.9. Comparaison des stratégies de fusion sémantique (BERT 2 couches + DenseNet Bloc 4)

Stratégie de Fusion	Préc.	Rapp.	Micro F1	Macro F1	Weighted F1	Kappa moy.
Concaténation simple	0,920	0,900	0,910	0,900	0,910	0,883
Low-Rank Fusion	0,920	0,900	0,910	0,890	0,910	0,877
Bilinear Fusion	0,910	0,910	0,910	0,890	0,910	0,877

L’analyse du Tableau 3.9 démontre que les trois mécanismes atteignent des performances globalement équivalentes, se stabilisant autour d’un Micro F1-score de 0,91. La concaténation simple affiche un Kappa moyen légèrement supérieur (0,883) face à la *Low-Rank Fusion* et à la *Bilinear Fusion* (toutes deux à 0,877), reflétant une meilleure concordance diagnostique réelle. Face à notre corpus de données médicales, la haute complexité paramétrique induite par des projections bilinéaires ou des espaces tensoriels de bas rang n’apporte aucun gain discriminatoire mesurable. Par conséquent, la concaténation simple est validée et conservée pour la suite de nos travaux en raison de sa grande stabilité algorithmique, de son immunité face au surapprentissage et de son efficience computationnelle.

L’analyse par label révèle que la concaténation simple surpasse ou égale les deux autres stratégies sur 8 labels sur 12. Les seuls labels où Low-Rank ou Bilinear obtiennent un Kappa légèrement supérieur sont Devices/Post-surgical (+0,042 pour Low-Rank : 0,790 vs 0,749) et Pleural Abnormality (+0,038 pour Bilinear : 0,869 vs 0,832), ce qui s’explique davantage par la nature ambiguë de ces classes que par une réelle supériorité de la méthode de fusion. En revanche, la concaténation simple obtient un Kappa nettement supérieur sur Diaphragmatic/Hiatal (0,981 vs 0,962 pour Low-Rank) et Cardiac Failure/Enlargement (0,860 vs 0,839 pour Bilinear). Ces observations suggèrent que, dans un contexte de données modérément volumineuses ($\approx 3\,000$ paires image-texte), les interactions inter-modalités capturées par des mécanismes plus complexes n’apportent pas de gain mesurable. La concaténation simple présente par ailleurs l’avantage d’une stabilité d’entraînement supérieure et d’une absence de paramètres supplémentaires susceptibles d’induire un surapprentissage.

la concaténation simple est compétitive ici pour :

- Faible risque de surapprentissage
- Absence de paramètres supplémentaires de fusion
- Meilleure généralisation quand le nombre d'exemples est limité

3.5.3 Analyse comparative des stratégies de fine-tuning

3.5.3.1 Impact du fine-tuning de BioClinicalBERT

Trois stratégies de fine-tuning ont été envisagées et comparées dans le cadre de ce travail, correspondant à trois niveaux croissants d'adaptation du modèle à la tâche de classification thoracique multi-label.

Stratégie 1 — Encodeur entièrement gelé (sans fine-tuning) : tous les paramètres de BioClinicalBERT sont gelés. Le modèle est utilisé exclusivement comme extracteur de caractéristiques statiques, sans aucune adaptation à la tâche. Les résultats obtenus illustrent concrètement les limites de cette approche :

- Micro F1 = 0,47
- Kappa moyen = 0,317 (accord qualifié de faible à acceptable selon l'échelle de Landis & Koch)

On observe notamment des effondrements de précision sur des classes diagnostiquement importantes :

- Mass / Nodule : précision de 0,087 (95 faux positifs)
- Diaphragmatic / Hiatal : précision de 0,258

L'instabilité globale des prédictions se traduit par une multiplication des faux positifs sur l'ensemble des classes. Ces résultats confirment qu'un encodeur textuel non adapté au domaine introduit du bruit sémantique plutôt qu'un signal discriminant, rendant la fusion multimodale inefficace.

Stratégie 2 — Dégel de la dernière couche uniquement : cette configuration libère un nombre de couches intermédiaire entre la stratégie figée et la stratégie profonde. Les résultats obtenus montrent une nette amélioration par rapport au modèle figé, avec des performances solides sur la plupart des classes :

- Micro F1 = 0,89
- Kappa moyen = 0,850 (accord presque parfait pour 9 classes, substantiel pour 3 classes)

Les scores par classe sont homogènes et élevés :

- Aortic / Vascular Change : $F1 = 0,95$ (précision 0,95, rappel 0,95)
- Diaphragmatic / Hiatal : $F1 = 0,98$ (rappel parfait à 1,00)
- Normal : $F1 = 0,97$
- Granulomatous / Fibrotic Scar : $F1 = 0,90$

Les rares difficultés persistent sur des classes à faible support (Mass / Nodule : $F1 = 0,73$, Skeletal / Thoracic Deformity : $F1 = 0,78$), mais sans effondrement des métriques. Cette stratégie constitue donc un compromis efficace entre adaptation et préservation des connaissances pré-entraînées.

Stratégie 3 — Dégel des deux dernières couches et du pooler (configuration retenue) : seules les couches 11 et 12 de BioClinicalBERT, ainsi que le pooler, sont rendues entraînaibles. Les résultats obtenus sont nettement supérieurs à ceux de l’encodeur entièrement gelé et légèrement supérieurs à ceux de la stratégie 2 :

- Micro $F1 = 0,91$
- Kappa moyen = 0,883 (accord presque parfait pour 10 classes, substantiel pour 2 classes)

L’écart de Micro $F1$ entre la stratégie figée et cette stratégie ($0,47 \rightarrow 0,91$) représente un gain absolu de 0,44 points. Par rapport à la stratégie 2, le gain est plus modeste mais non négligeable (+0,02 en Micro $F1$, +0,033 en Kappa moyen), avec une amélioration particulièrement nette sur les classes les plus difficiles (Mass / Nodule : $F1$ passe de 0,73 à 0,87 ; Pleural Abnormality conserve un $F1$ -score stable à 0,85, avec un rappel plus élevé dans la stratégie à deux couches, au prix d’une précision plus faible). Cette stratégie valide sans ambiguïté la supériorité d’un fine-tuning ciblé sur les couches les plus profondes dans notre contexte expérimental.

justification du choix de la stratégie : Dans un modèle comme BioClinicalBERT (12 couches Transformer), toutes les couches ne jouent pas le même rôle. Les couches inférieures (1 à 4) capturent des informations syntaxiques et lexicales très générales, peu spécifiques à un domaine ou une tâche. Les couches moyennes (5 à 8) encodent des relations sémantiques plus larges. En revanche, les couches supérieures (9 à 12) et le pooler stockent les représentations les plus proches de la tâche et du domaine de pré-entraînement (ici, les notes cliniques MIMIC-III).

Dégeler uniquement les deux dernières couches (11 et 12) ainsi que le pooler présente trois avantages décisifs :

1. **Adaptation efficace à la tâche :**

Ces couches sont les plus à même d’ajuster leurs représentations à la classification multi-label des rapports thoraciques, sans altérer les compétences linguistiques générales acquises par les couches inférieures.

2. **Prévention du catastrophic forgetting :**

Un fine-tuning trop profond (dégeler toutes les couches) risquerait de détruire les représentations cliniques riches apprises lors du pré-entraînement sur MIMIC-III, surtout avec un dataset de taille modérée. En limitant le nombre de couches entraînaibles, on préserve ces connaissances tout en les adaptant légèrement.

3. **Réduction drastique du nombre de paramètres entraînaibles :**

Seuls environ 14 millions de paramètres sont réentraînés (contre 110 millions pour le modèle complet), soit une réduction de 87 % du budget paramétrique textuel. Cela limite le risque de surapprentissage et accélère l’entraînement.

3.5.3.2 Impact du fine-tuning de DenseNet121

Afin d’évaluer la contribution de la modalité visuelle de manière isolée puis conjointe, deux configurations supplémentaires ont été testées en maintenant BioClinicalBERT dégelé sur ses deux dernières couches (Tests 4 et 5 du Tableau 3.6).

Configuration 1 — DenseNet121 entièrement gelé (Test 4) :

L’évaluation du modèle combinant un BERT optimisé à un DenseNet121 intégralement gelé affiche des scores compétitifs :

- Micro F1 = 0,920
- Macro F1 = 0,900
- Kappa moyen = 0,883

Ces performances montrent que les caractéristiques visuelles génériques issues du pré-entraînement ImageNet suffisent à maintenir un niveau de performance global élevé. Toutefois, en restant gelé, l’extracteur visuel produit des descripteurs génériques du domaine général, limitant la capacité du modèle à capturer des caractéristiques pathologiques fines spécifiques au domaine médical.

Configuration 2 — Dégel des denseblocks 3 et 4 (Test 5, configuration retenue) :

Le dégel conjoint des blocs denseblock3 et denseblock4 permet d’élever le Kappa moyen à sa valeur maximale absolue :

- Micro F1 = 0.915
- Macro F1 = 0.900
- Kappa moyen = 0.887

Le coefficient Kappa, qui mesure le degré de concordance réel en corrigeant l'effet du hasard, valide la supériorité clinique de cette configuration par rapport au DenseNet entièrement gelé.

TAB. 3.10. Analyse incrémentale des stratégies de fine-tuning de l'architecture unifiée

Test	Configuration BioClinical-BERT	Configuration DenseNet121	Micro F1	Macro F1	Kappa moy.
<i>Phase 1 : Optimisation de la modalité textuelle (DenseNet Bloc 4 dégelé)</i>					
1	Entièrement gelé	Dégel du <i>dense-block4</i>	0,470	0,410	0,317
2	Dégel de la dernière couche	Dégel du <i>dense-block4</i>	0,890	0,870	0,850
3	Dégel des 2 dernières couches	Dégel du <i>dense-block4</i>	0,910	0,900	0,883
<i>Phase 2 : Optimisation de la modalité visuelle (BERT 2 couches dégelées)</i>					
4	Dégel des 2 dernières couches	Entièrement gelé	0,920	0,900	0,883
5	Dégel des 2 dernières couches	Dégel des <i>dense-blocks</i> 3 et 4	0,915	0,900	0,887

Justification du choix de la configuration :

Dans l'architecture DenseNet121, les deux premiers blocs denses (denseblock1 et denseblock2) capturent des caractéristiques génériques de bas niveau (contours, textures), tandis que denseblock3 et denseblock4 encodent des représentations sémantiques de haut niveau spécifiques au contenu de l'image. Dégeler ces deux derniers blocs présente trois avantages :

- **Spécialisation au domaine médical** : les blocs profonds peuvent s'adapter aux caractéristiques radiologiques spécifiques (opacités, nodules, déformations osseuses) non présentes dans ImageNet.
- **Préservation des filtres génériques** : les deux premiers blocs restent gelés, conservant les détecteurs de bas niveau robustes et évitant le surapprentissage sur un dataset de taille modérée.
- **Équilibre paramétrique avec BioClinicalBERT** : cette configuration aligne la capacité d'adaptation visuelle avec la richesse sémantique issue des couches profondes

de BioClinicalBERT, garantissant qu’aucune modalité ne domine l’autre lors de la fusion.

Ce choix s’aligne symétriquement avec celui fait pour BioClinicalBERT (dégel des deux dernières couches), garantissant un équilibre entre les deux modalités visuelle et textuelle au sein du pipeline multimodal final.

3.5.4 Limites de l’approche

- Extracteurs non conjointement affinés, avec un potentiel gain dans une approche end-to-end
- Fusion simple ne modélise pas les interactions fines entre modalités
- Dataset mono institutionnel

3.5.5 Pistes d’amélioration

- Réévaluer les fusions avancées (Low Rank, Bilinear) sur un dataset plus large
- Fine tuning conjoint vision-langage
- Ajout de mécanismes d’attention croisée entre les modalités

3.6 Interface de démonstration : Gradio

Afin de valoriser notre approche et de rendre le modèle multimodal accessible, interactif et testable en temps réel, nous avons développé une interface de démonstration en utilisant la bibliothèque **Gradio**. Cet outil open-source en Python permet de générer rapidement des applications web légères, idéales pour simuler un cas d’usage en environnement clinique.

L’implémentation de cette interface répond à un double objectif scientifique et applicatif :

- **Simulation du parcours clinique** : Elle reproduit la démarche d’un clinicien en lui permettant de charger simultanément les deux incidences radiographiques et de renseigner le compte rendu textuel associé.
- **Aide à la décision médicale** : Elle fournit instantanément une évaluation prédictive automatique, affichant clairement la distribution des probabilités et les alertes pour chaque pathologie détectée.

3.6.1 Architecture logicielle et flux de traitement

L'interface a été structurée de manière intuitive pour suivre le flux de travail du personnel médical, se divisant en deux modules principaux :

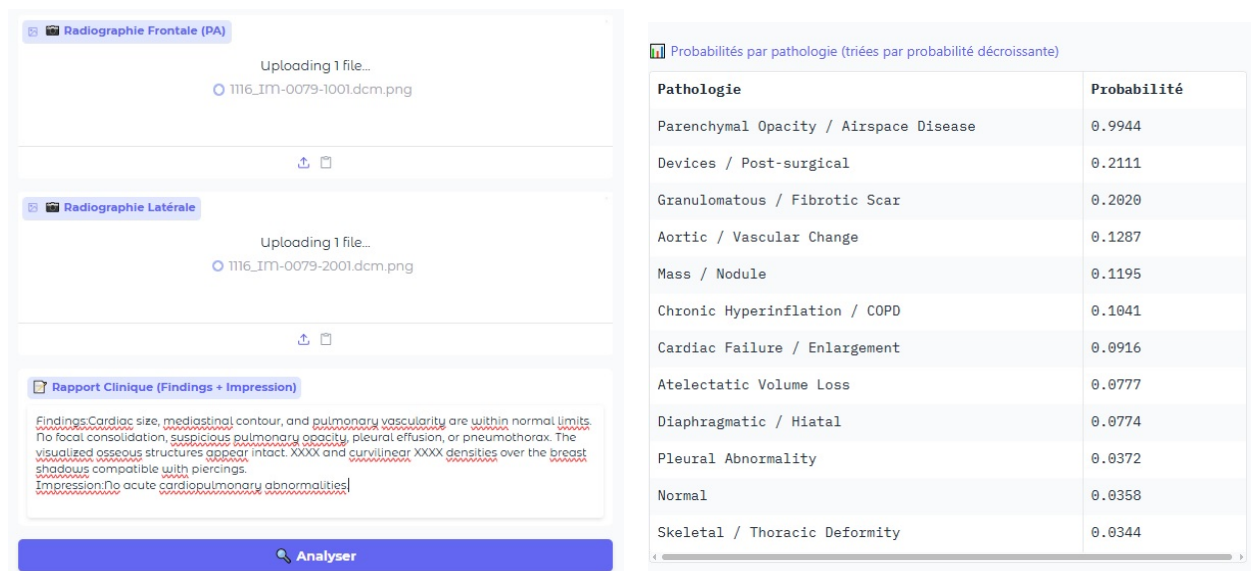
1. Module d'entrée et de saisie clinique : Comme mis en évidence par la figure 3.10a, l'utilisateur dispose d'un espace de chargement double acceptant les deux vues radiographiques (frontale et latérale). Cet espace est complété par un champ de texte libre dédié aux sections *Findings* et *Impression* du rapport médical. Un bouton unique déclenche l'analyse en transmettant ces données hétérogènes au pipeline multimodal.

2. Restitution globale des résultats : Une fois l'inférence terminée, l'interface affiche l'ensemble des résultats de manière centralisée (figure 3.10b). À droite des images, un tableau dynamique intègre des jauges visuelles colorées représentant les scores de confiance pour les pathologies pulmonaires. En bas de page, un rapport textuel synthétique est produit automatiquement pour résumer le statut de chaque classe diagnostique. L'ensemble de ces données reste sélectionnable afin d'en faciliter la copie.

3.6.2 Illustration de l'interface sur un cas réel

Les figures ci-dessous illustrent le comportement complet de l'application suite à l'analyse d'un cas clinique réel issu de la base de données *Indiana University Chest X-ray*.

La zone de saisie, détaillée dans la Figure 3.10a, permet de charger et de configurer simultanément les examens d'imagerie et le rapport textuel initial.



(a) Configuration de la zone de saisie (images et texte).

(b) Vue globale des prédictions et du rapport généré.

FIG. 3.10. Démonstration de l'interface Gradio lors de l'évaluation d'un cas multimodal.

L'analyse de cet exemple, dont la vue globale des résultats est exposée dans la Figure 3.10b, montre que le modèle fusionne efficacement ces informations. Il identifie ainsi correctement la pathologie *Parenchymal Opacity / Airspace Disease* avec une probabilité très élevée (0,9944), clairement visible sur le panneau des prédictions de la Figure 3.10b. Les autres pathologies présentent des probabilités faibles, ce qui indique une prédiction cohérente et focalisée sur l'anomalie principale détectée dans les images et mentionnée dans le rapport clinique.

L'application remplit ainsi parfaitement son rôle de preuve de concept, offrant une vitrine concrète et fonctionnelle de l'approche multimodale développée dans ce travail de recherche.

3.7 Conclusion

Nous avons présenté une approche multimodale robuste pour la classification multi-label de radiographies thoraciques, basée sur une concaténation simple entre caractéristiques DenseNet121 et BioClinicalBERT. Les résultats montrent l'efficacité de cette architecture sur le dataset Indiana, avec un Micro F1 de 0.91. La concaténation simple, bien que moins sophistiquée que des méthodes de fusion avancées, s'avère ici particulièrement adaptée compte tenu de la taille modérée du jeu de données et de la volonté de limiter le surapprentissage. Ce chapitre pose ainsi une base solide pour d'éventuelles améliorations futures (fusion bilinéaire, fine-tuning conjoint, etc.).

Conclusion générale

L'objectif de ce mémoire était de développer un modèle intelligent pour le diagnostic médical, basé sur une approche multimodale combinant l'analyse d'images radiographiques et le traitement de rapports médicaux textuels. Face aux limites des approches unimodales qui exploitent uniquement l'image ou uniquement le texte, nous avons proposé une architecture fusionnant un réseau de neurones convolutif (DenseNet121) pour l'extraction des caractéristiques visuelles et un modèle de langage spécialisé (BioClinicalBERT) pour l'analyse sémantique des comptes rendus.

Nos expérimentations menées sur le jeu de données Indiana University Chest X-ray, après un regroupement rigoureux des 118 labels initiaux en 12 catégories clinico-radiologiques, ont montré des résultats prometteurs. Le modèle multimodal par concaténation simple a atteint un Micro F1-score de 0,91 et un Kappa moyen de 0,887, démontrant une excellente capacité de classification multi-label des pathologies thoraciques. L'analyse par groupe diagnostique a révélé des performances élevées pour des classes telles que Aortic/Vascular Change ($F1=0,97$) et Diaphragmatic/Hiatal ($F1=0,97$), tout en identifiant des marges d'amélioration pour des catégories plus hétérogènes comme Devices/Post-surgical et Skeletal/Thoracic Deformity.

La comparaison avec d'autres stratégies de fusion (Low-Rank et Bilinear) a confirmé que, dans un contexte de données modérément volumineuses (environ 3 000 paires image-texte), la concaténation simple constitue un choix judicieux, offrant stabilité, interprétabilité et un faible risque de surapprentissage.

Ce travail ouvre plusieurs perspectives. Sur le plan méthodologique, l'intégration d'un fine tuning conjoint des extracteurs vision et langage, ainsi que l'ajout de mécanismes d'attention croisée entre modalités, pourraient améliorer la modélisation des interactions fines. Sur le plan applicatif, l'évaluation sur des datasets multicentriques et plus volumineux permettrait de tester la généralisabilité du modèle. Enfin, une validation clinique auprès de radiologues serait essentielle pour évaluer l'acceptabilité et l'utilité réelle d'un tel outil d'aide à la décision en pratique médicale.

En conclusion, ce mémoire contribue à démontrer la pertinence des approches multimodales pour le diagnostic médical et propose une base solide pour de futures recherches visant à rapprocher l'intelligence artificielle du raisonnement clinique humain.

Synthèse de la littérature : Analyse des articles

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
1	Andre Esteva et al. (2019)	A guide to deep learning in healthcare / Nature Medicine	Deep Learning, IA en santé, Réseaux neuronaux	Article de revue	Hautes performances en imagerie médicale, prédiction de maladies	Aperçu complet des applications du deep learning en santé	Qualité des données, interprétabilité, vie privée, intégration clinique
2	Geert Litjens et al. (2017)	A survey on deep learning in medical image analysis / Medical Image Analysis	Deep Learning, CNN, Vision par ordinateur	Article de revue	Amélioration significative en classification, segmentation, détection d'images médicales	Première revue complète sur le deep learning en imagerie médicale	Manque de données annotées, coûts computationnels, interprétabilité limitée
3	Alex Krizhevsky et al. (2012)	ImageNet Classification with Deep Convolutional Neural Networks / NIPS	Deep Learning, CNN, Classification d'images	Recherche expérimentale	Performance état-de-l'art sur ImageNet avec AlexNet	Architecture fondatrice qui a popularisé le deep learning en vision	Coût computationnel, besoin de grandes données, sur-apprentissage
4	Ashish Vaswani et al. (2017)	Attention Is All You Need / NIPS	Architecture Transformer, Mécanisme d'attention	Étude expérimentale	Supériorité sur les modèles séquence-à-séquence traditionnels	Introduction du modèle Transformer, base de BERT, GPT	Exigences mémoire et computationnelles pour les grands modèles
5	Jacob Devlin et al. (2019)	BERT : Pre-training of Deep Bidirectional Transformers / NAACL-HLT	Transformers bidirectionnels, TAL, Transfer learning	Étude expérimentale	Résultats état-de-l'art sur benchmarks NLP	Révolution du TAL par pré-entraînement bidirectionnel	Coût computationnel élevé, adaptation de domaine, biais
6	Kexin Huang et al. (2019)	ClinicalBERT : Modeling Clinical Notes and Predicting Hospital Readmission / arXiv	TAL clinique, Transformers bidirectionnels, Transfer learning	Étude expérimentale	Amélioration de la prédiction de réadmission hospitalière	Adaptation de BERT au domaine clinique	Disponibilité limitée des données cliniques annotées, vie privée
Suite sur page suivante...							

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
7	Farah Khader et al. (2023)	Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters / Radiology	Deep Learning multimodal, Transformers, Fusion de données	Étude expérimentale	Melleure performance avec fusion radiographies + données cliniques	Démontre l'efficacité des transformers multimodaux en santé	Besoin de datasets multicentriques, validation externe, intégration clinique
8	Tadas Baltrušaitis et al. (2019)	Multimodal Machine Learning : A Survey and Taxonomy / IEEE TPAMI	Apprentissage machine multimodal, Fusion, Représentation	Article de revue	Amélioration des performances par fusion de modalités complémentaires	Taxonomie structurée de référence pour l'apprentissage multimodal	Robustesse, modèles explicables, gestion des modalités manquantes
9	Eric J. Topol (2019)	High-performance medicine : The convergence of human and artificial intelligence / Nature Medicine	Intelligence artificielle, Médecine numérique, Apprentissage profond	Article de revue	Amélioration du diagnostic, de la prise de décision clinique et de la médecine personnalisée	Présente une vision globale de l'intégration de l'IA en médecine et de son potentiel pour améliorer les soins	Qualité des données, applicabilité des modèles, intégration dans la pratique clinique, questions éthiques
10	James H. Thrall et al. (2018)	Artificial Intelligence and Machine Learning in Radiology : Opportunities, Challenges, Pitfalls, and Criteria for Success / Journal of the American College of Radiology	Intelligence artificielle, Apprentissage automatique, Radiologie	Article de revue	Mise en évidence des opportunités de l'IA pour améliorer le diagnostic, le flux de travail et l'interprétation des images médicales	Analyse les opportunités, les défis et les conditions nécessaires à l'intégration réussie de l'IA en radiologie	Disponibilité et qualité des données, validation clinique, explicabilité des modèles et intégration dans les services de radiologie
11	A. Dali (2022)	Segmentation à base d'atlas d'images médicales / Thèse	Segmentation d'images médicales, Apprentissage en ligne	Recherche doctorale	Amélioration de l'adaptabilité et précision de segmentation	Avancement des techniques de segmentation adaptative	Complexité computationnelle, variabilité des datasets
12	S. Mithun et al. (2023)	Clinical Concept-Based Radiology Reports Classification / JDI	TAL clinique, Classification de rapports radiologiques	Étude expérimentale	Amélioration de la classification des rapports de carcinome pulmonaire	Pipeline de classification automatique pour rapports radiologiques	Taille des datasets, variabilité des structures de rapports
13	Rafael C. Gonzalez & Richard E. Woods (2018)	Digital Image Processing / Pearson	Traitement d'images numériques, Vision par ordinateur	Manuel académique	Techniques d'amélioration, segmentation, extraction de caractéristiques	Référence fondamentale en traitement d'images numériques	Intégration du ML/DL avec méthodes traditionnelles

Suite sur page suivante...

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
22	L'équipe de développement de scikit-image (2026)	Image Denoising	Filtrage du bruit, Variation Totale (TV), Ondelettes	Documentation et exemples de code	Restauration de la fidélité de l'image par atténuation des composantes hautes fréquences	Banc d'essai comparatif d'algorithmes de lissage préservant les contours	Nécessite un ajustement manuel des hyperparamètres selon le type de bruit
23	Adrien Depeursinge et al. (2017)	Fundamentals of Texture Processing / Biomedical Texture Analysis	Analyse de texture biomédicale, Matrices GLCM, Filtrés Gabor	Chapitre d'ouvrage / Revue théorique	Extraction robuste de descripteurs de textures pathologiques	Base algorithmique pour le traitement de textures en imagerie médicale	Sensibilité au bruit, coût computationnel sur de gros volumes de données
24	L'équipe de développement d'OpenCV (2026)	Histogram Equalization	Traitement d'images temps réel, Transformation d'intensité	Documentation technique logicielle	Optimisation des calculs de distribution de niveaux de gris en C++/Python	Outils d'évaluation rapide et d'affichage de distributions de pixels	Orienté vers l'efficacité d'exécution plus que vers la modélisation théorique
25	IAEA (2020)	Image Quality and Quality Control in Diagnostic Nuclear Medicine / IAEA	Qualité d'image, Médecine nucléaire, Contrôle qualité	Guide technique	Cadre pour l'optimisation de la qualité d'image en médecine nucléaire	Référence pour le contrôle qualité en imagerie nucléaire	Automatisation QC avancée, évaluation qualité basée sur IA
26	Andreas Zwanenburg et al. (2020)	The Image Biomarker Standardization Initiative / Radiology	Radiomique, Biomarqueurs d'image, Standardisation	Étude collaborative	Standardisation des définitions des caractéristiques radiomiques	Initiative IBSI pour la reproductibilité en radiomique	Expansion des caractéristiques standardisées, validation clinique
27	ISO/IEC (1994)	ISO/IEC 10918-1:1994 JPEG Standard	Compression d'image, JPEG, Compression avec perte	Norme technique	Compression efficace d'images avec qualité visuelle acceptable	Établissement du standard JPEG mondialement adopté	Efficacité de compression améliorée, méthodes sans perte
28	W3C (2003)	Portable Network Graphics (PNG) Specification / ISO/IEC	Compression d'image, Compression sans perte	Spécification technique	Compression sans perte efficace avec support de transparence	Standard PNG pour graphiques web et imagerie numérique	Optimisation pour systèmes de traitement d'images grande échelle
29	CompuServe (1990)	Graphics Interchange Format, Version 89a	Encodage d'image, Compression sans perte	Spécification technique	Stockage efficace d'images basse couleur, animation	Format GIF89a pionnier pour graphiques web	Profondeur couleur limitée (256 couleurs), efficacité limitée
30	Microsoft (2021)	BMP Format Overview / Microsoft Learn	Imagerie bitmap, Stockage d'images	Documentation technique	Stockage simple et direct d'images raster haute fidélité	Format BMP fondamental pour systèmes Windows	Tailles de fichiers importantes, compression limitée
Suite sur page suivante.							

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
40	Duan et al. (2024)	Deep learning based multimodal biomedical data fusion / Information Fusion	Fusion multimodale, Deep learning	Revue comparative	Le DL améliore significativement les performances diagnostiques	Taxonomie unifiée et cadre d'évaluation pour la fusion biomédicale	Métriques standardisées, modèles légers pour déploiement clinique
41	Samuel, A. L. (1959)	Some studies in machine learning using the game of checkers / IBM JRD	Apprentissage par renforcement	Conception expérimentale	Programme atteignant niveau respectable au jeu de dames	Première démonstration d'un programme apprenant par l'expérience	Ressources computationnelles limitées, features sélectionnées manuellement
42	LeCun, Bengio & Hinton (2015)	Deep learning / Nature	Réseaux neuronaux profonds, Représentation	Revue conceptuelle	Apprentissage de représentations hiérarchiques à partir de données brutes	Synthèse de référence ayant popularisé le deep learning	Compréhension théorique de la généralisation, besoin d'apprentissage non supervisé
43	University of Hull (2021)	Machine Learning vs Deep Learning / HPC MediaWiki	Paradigmes de l'apprentissage automatique et profond	Documentation et tutoriel technique	Clarification des frontières conceptuelles et des prérequis matériels (HPC/GPU)	Guide de vulgarisation structuré pour les infrastructures de calcul	Évolution rapide des frameworks par rapport aux plateformes figées
44	LeCun et al. (1998)	Gradient-based learning applied to document recognition / Proc. IEEE	CNN (LeNet), Rétropropagation	Développement et évaluation	LeNet-5 atteint 0.8% d'erreur sur MNIST	Article fondateur introduisant les CNN modernes	Images limitées (32x32), architectures peu profondes
45	Rikiya Yamashita et al. (2018)	Convolutional neural networks : an overview in radiology / Insights into Imaging	Fondations des CNN, Extraction de caractéristiques, Radiologie	Article de revue didactique	Synthèse accessible des concepts de convolution, pooling et backpropagation	Vulgarisation et guide d'application des réseaux de neurones pour les radiologues	Manque d'approches multimodales intégrées (texte + image)
46	Simonyan & Zisserman (2014)	Very deep convolutional networks / arXiv	CNN profonds, Petits filtres	Évaluation expérimentale	La profondeur accrue améliore la précision jusqu'à 16-19 couches	Introduction de VGGNet, principe de conception par piles de petits filtres	Coût computationnel élevé, techniques pour réseaux plus profonds
47	He et al. (2016)	Deep residual learning for image recognition / CVPR	Apprentissage résiduel, Connexions shortcut	Conception expérimentale	ResNet-152 couches : 3.57% d'erreur top-5 sur ImageNet	Invention de l'apprentissage résiduel, permet les réseaux extrêmement profonds	Compréhension théorique des connexions résiduelles
48	Huang et al. (2017)	Densely connected convolutional networks / CVPR	Connectivité dense, Réutilisation des caractéristiques	Évaluation expérimentale	DenseNet : état-de-l'art avec moins de paramètres que ResNet	Introduction de DenseNet, connectivité dense comme alternative	Consommation mémoire à l'entraînement, compréhension théorique
							Suite sur page suivante...

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
49	Tan & Le (2019)	EfficientNet : Rethinking model scaling for CNNs / ICML	Mise à l'échelle composée, NAS	Étude systématique	EfficientNet-B7 : 84.4% accuracy avec 8.4x moins de paramètres	Introduction du principe de mise à l'échelle composée	Hypothèses de relations monotones, NAS non pleinement optimisé
50	Rojas-Carabali et al. (2024)	Natural Language Processing in medicine and ophthalmology : A review for the 21st-century clinician / Asia-Pacific Journal of Ophthalmology	Traitement automatique du langage médical, Représentations textuelles (Bag-of-Words, TF-IDF, embeddings)	Article de revue	Synthèse des techniques de TAL appliquées au domaine clinique et ophtalmologique	Vue d'ensemble pédagogique des méthodes de représentation textuelle en NLP médical	Hétérogénéité des corpus cliniques, besoin de modèles spécialisés par spécialité
51	Lee et al. (2020)	BioBERT : a pre-trained biomedical language model / Bioinformatics	Pré-entraînement spécifique au domaine	Pré-entraînement BERT	BioBERT surpasse BERT sur NER, RE, QA biomédicaux	Modèle de langue biomédical fondateur pour TAL santé	Texte seulement, coût computationnel, anglais uniquement
52	Alsentzer et al. (2019)	Publicly available clinical BERT embeddings / Clinical NLP Workshop	Transfer learning domaine clinique	Libération de modèles	Amélioration des performances sur tâches TAL clinique	Embeddings BERT cliniques librement accessibles	Basé sur MIMIC-III anglais seulement, multimodalité limitée
53	Yang et al. (2022)	A large language model for electronic health records / npj Digital Med.	Grands modèles de langue cliniques	Entraînement de GatorTron	GatorTron surpasse les modèles existants sur tâches TAL cliniques	GatorTron, un des plus grands modèles de langue cliniques	Basé sur un seul système de santé, validation externe nécessaire
54	Benomar, M. L. & Gherbi, D. (à paraître)	Comparative Analysis of Vector-Level Multimodal Fusion Strategies / À paraître	Fusion multimodale au niveau vecteur	Conception expérimentale	À compléter selon article original	Cadre comparatif pour stratégies de fusion multimodale	Diversité limitée des données, gestion des modalités manquantes
55	Md Ashik Khan & Md Nahid Siddique (2025)	Fixed-Budget Parameter-Efficient Training with Frozen Encoders Improves Multimodal Chest X-Ray Classification / ICCIT	Entraînement parameter-efficient (LoRA, BitFit, Encoders, Adapters), Encoders gelés, Fusion multimodale	Étude expérimentale contrôlée sur Indiana University Chest X-Ray	Les méthodes PET (0,89-0,91 AUROC) surpassent le finetuning complet (0,77 AUROC) avec 40x moins de paramètres ; gain multimodal limité avec un encodeur textuel généraliste (DistilBERT)	Démontre que les gains multimodaux proviennent surtout de l'allocation des paramètres plutôt que d'une réelle synergie inter-modale avec un encodeur texte générique	Calibration dégradée (ECE élevé), nécessité d'encoders biomédicaux spécialisés (BioBERT, ClinicalBERT) pour renforcer la synergie
							Suite sur page suivante...

N°	Auteurs / Année	Titre / Journal	Théories	Méthodologie	Résultats	Contribution significative	Recommandations / Gaps
56	Mojtaba Jahanian et al. (2025)	Multimodal transformers for joint diagnosis and retrieval from chest X-rays and radiology reports / Informatics in Medicine Unlocked	Fusion par attention croisée bidirectionnelle profonde, Apprentissage contrastif (InfoNCE), Vision Transformer + ClinicalBERT	Étude expérimentale sur MIMIC-CXR et IU X-ray, avec étude d'ablation	AUC de 0,847 sur MIMIC-CXR, surpassant les baselines unimodales et multimodales superflues ; le retrait de ClinicalBERT cause la plus forte dégradation de performance	Modèle RadFuse : architecture unifiée combinant classification multi-label et recherche image-rapport via fusion profonde token-level	Coût computationnel élevé, dépendance à un alignement parfait image-rapport, généralisation à d'autres modalités d'imagerie non testée
57	National Library of Medicine (2026)	OpenI : Open Access Biomedical Image Search Engine	Moteurs de recherche biomédicaux, Accès ouvert	Plateforme technologique	Indexation massive de cas cliniques et imageries associées	Infrastructure majeure pour le test d'algorithmes de vision-langage	Hétérogénéité des formats de données et résolutions d'images
58	R. Chetan et al. (2023)	HybridTransferNet : Advancing Soil Image Classification / Research Square	Apprentissage par transfert hybride, Feature fusion	Modélisation expérimentale	Validation de l'agrégation de caractéristiques issues de backbones distincts	Preuve de concept sur l'efficacité des architectures de transfert hybrides	Évaluation limitée à un domaine d'application spécifique (sols)

Bibliographie

- [1] Andre Esteva, Alex Robicquet, Bharath Ramsundar, et al. A guide to deep learning in healthcare. *Nature Medicine*, 2019.
- [2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT : Modeling clinical notes and predicting hospital readmission. *arXiv preprint*, 2019.
- [7] Farah Khader, Gregor Müller-Franzes, Tianyu Wang, et al. Multimodal deep learning for integrating chest radiographs and clinical parameters : A case for transformers. *Radiology*, 2023.
- [8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning : A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [9] Eric J. Topol. High-performance medicine : The convergence of human and artificial intelligence. *Nature Medicine*, 25(1) :44–56, 2019.

- [10] James H. Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, and James A. Brink. Artificial intelligence and machine learning in radiology : Opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, 15(3) :504–508, 2018.
- [11] A. Dali. *Segmentation à base d’atlas d’images médicales utilisant l’apprentissage en ligne*. PhD thesis, École nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2022.
- [12] S. Mithun, A. K. Jha, U. B. Sherkhane, et al. Clinical concept-based radiology reports classification pipeline for lung carcinoma. *Journal of Digital Imaging*, 2023.
- [13] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.
- [14] Wilhelm Burger and Mark J. Burge. *Digital Image Processing : An Algorithmic Introduction*. Springer, 2022.
- [15] SSTC – Politique scientifique fédérale belge. Les images numériques. <https://eo.belspo.be/fr/les-images-numeriques>. Consulté le 12 juin 2026.
- [16] Fabio Medda. Comprendre la résolution d’image : Dpi et ppi. <https://fmedda.com/fr/article/dpi-ppi>. Consulté le 12 juin 2026.
- [17] Contributeurs de Wikipédia. Fichier : Gradient2.svg. <https://en.wikipedia.org/wiki/File:Gradient2.svg>. Consulté le 12 juin 2026.
- [18] Eli Peli. Contrast in complex images. *Journal of the Optical Society of America A*, 1990.
- [19] L’équipe de développement de scikit-image. Histogram equalization. https://scikit-image.org/docs/stable/auto_examples/color_exposure/plot_equalize.html. Consulté le 12 juin 2026.
- [20] Lesia Mochurad. Approach for enhancing the accuracy of semantic segmentation of chest x-ray images by edge detection and deep learning integration. *Frontiers in Artificial Intelligence*, 2025.
- [21] James C. Ross, Raúl San José Estépar, Alejandro Díaz, Carl-Fredrik Westin, Ron Kikinis, Edwin K. Silverman, and George R. Washko. Lung extraction, lobe segmentation and hierarchical region assessment for quantitative analysis on high resolution computed tomography images. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2009.

- [22] L'équipe de développement de scikit-image. Image denoising. https://scikit-image.org/docs/stable/auto_examples/filters/plot_denoise.html. Consulté le 12 juin 2026.
- [23] Adrien Depeursinge, Julien Fageot, and Omar S. Al-Kadi. Fundamentals of texture processing for biomedical image analysis. *Biomedical Texture Analysis : Fundamentals, Tools and Challenges*, 2017.
- [24] L'équipe de développement d'OpenCV. Histogram equalization. https://docs.opencv.org/3.4/d4/d1b/tutorial_histogram_equalization.html. Consulté le 12 juin 2026.
- [25] International Atomic Energy Agency. Image quality and quality control in diagnostic nuclear medicine, 2020. <https://www.iaea.org/resources/rpop/health-professionals/nuclear-medicine/diagnostic-nuclear-medicine/image-quality-and-quality-control>.
- [26] Andreas Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, et al. The image biomarker standardization initiative : Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 2020.
- [27] ISO/IEC. Digital compression and coding of continuous-tone still images : Requirements and guidelines, 1994. ISO/IEC 10918-1:1994.
- [28] World Wide Web Consortium. Portable network graphics (png) specification (second edition), 2003. <https://www.w3.org/TR/PNG/>. Consulté le 12 juin 2026.
- [29] CompuServe. Graphics interchange format version 89a, 1990. <https://www.w3.org/Graphics/GIF/spec-gif89a.txt>. Consulté le 12 juin 2026.
- [30] Microsoft. BMP format overview, 2021. <https://learn.microsoft.com/en-us/windows/win32/wic/bmp-format-overview>.
- [31] Commissariat à l'énergie atomique et aux énergies alternatives. Livret pédagogique : Imagerie médicale, 2020. <https://www.cea.fr/multimedia/documents/publications/livrets-thematiques/livret-imagerie-medicaire.pdf>.
- [32] raddar. Chest x-rays (indiana university). <https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university>, 2020. Consulté en 2026.
- [33] DICOM Standards Committee. DICOM : Digital imaging and communications in medicine, 2024. <https://www.dicomstandard.org/about>.

- [34] A. P. Brady, B. Allen, J. Chong, E. Kotter, N. Kottler, J. Mongan, L. Oakden-Rayner, D. Pinto dos Santos, A. Tang, C. Wald, and J. Slavotinek. Developing, purchasing, implementing and monitoring ai tools in radiology : practical considerations. a multi-society statement from the acr, car, esr, ranzcr and rsna. *Radiology : Artificial Intelligence*, 6(1) :e230513, 2024.
- [35] Jean-Paul Beregi and François Le Jeune. Demande et compte-rendu des examens d'imagerie. Conseil National Professionnel de Médecine Nucléaire (CNP-MN), 2026. <https://www.cnp-mn.fr/les-fondamentaux-chapitre-13-demande-et-compte-rendu-des-examens-dimagerie/>. Consulté le 11 juin 2026.
- [36] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2) :304–310, 2016.
- [37] Zaheer Babar, Twan van Laarhoven, and Elena Marchiori. Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines. *PLOS ONE*, 17(1), 2022. Institute of Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands.
- [38] Stuart Russell and Peter Norvig. *Artificial Intelligence : A Modern Approach*. Pearson, 4th edition, 2021.
- [39] S. Kumar, S. Rani, S. Sharma, and H. Min. Multimodality fusion aspects of medical diagnosis : A comprehensive review. *Bioengineering*, 11(12) :1233, 2024.
- [40] J. Duan, J. Xiong, Y. Li, and W. Ding. Deep learning based multimodal biomedical data fusion : An overview and comparative review. *Information Fusion*, 112:102536, 2024.
- [41] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3) :210–229, 1959.
- [42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.
- [43] University of Hull. Programming : Deep learning - machine learning vs deep learning. <https://hpc.mediawiki.hull.ac.uk/>, 2021. Accédé en 2026.
- [44] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

- [45] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks : an overview and application in radiology. *Insights into Imaging*, 9(4) :611–629, aug 2018.
- [46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.
- [47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [48] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.
- [49] Mingxing Tan and Quoc V. Le. EfficientNet : Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019.
- [50] William Rojas-Carabali, Rajdeep Agrawal, Laura Gutierrez-Sinisterra, Sally L. Baxter, Carlos Cifuentes-González, Yap Chun Wei, John Abisheganaden, Palvannan Kannapiran, Sunny Wong, Bernett Lee, Alejandra de-la Torre, and Rupesh Agrawal. Natural language processing in medicine and ophthalmology : A review for the 21st-century clinician. *Asia-Pacific Journal of Ophthalmology*, 13(4) :100084, 2024. PMCID : PMC11919464.
- [51] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240, 2020.
- [52] Emily Alsentzer, John Murphy, William Boag, et al. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.
- [53] Xi Yang, Aokun Chen, Nima PourNejatian, et al. A large language model for electronic health records. *npj Digital Medicine*, 5:194, 2022.
- [54] M. L. Benomar and D. Gherbi. Comparative analysis of vector-level multimodal fusion strategies for medical diagnostic classification. *Journal to be confirmed*, 2026.
- [55] Md Ashik Khan and Md Nahid Siddique. Fixed-budget parameter-efficient training with frozen encoders improves multimodal chest x-ray classification. *2025 28th International Conference on Computer and Information Technology (ICCIT)*, 2025. arXiv :2512.21508.

-
- [56] Mojtaba Jahanian, Abbas Karimi, Nafiseh Osati Eraghi, and Faraneh Zarafshan. Multimodal transformers for joint diagnosis and retrieval from chest x-rays and radiology reports. *Informatics in Medicine Unlocked*, 58:101672, 2025.
 - [57] National Library of Medicine. Openi : Open access biomedical image search engine. <https://openi.nlm.nih.gov/>. Consulté en 2026.
 - [58] R. Chetan, D. V. Ashoka, and B. V. Ajay Prakash. Hybridtransfernet : Advancing soil image classification through comprehensive evaluation of hybrid transfer learning. *Research Square*, 2023. Preprint.