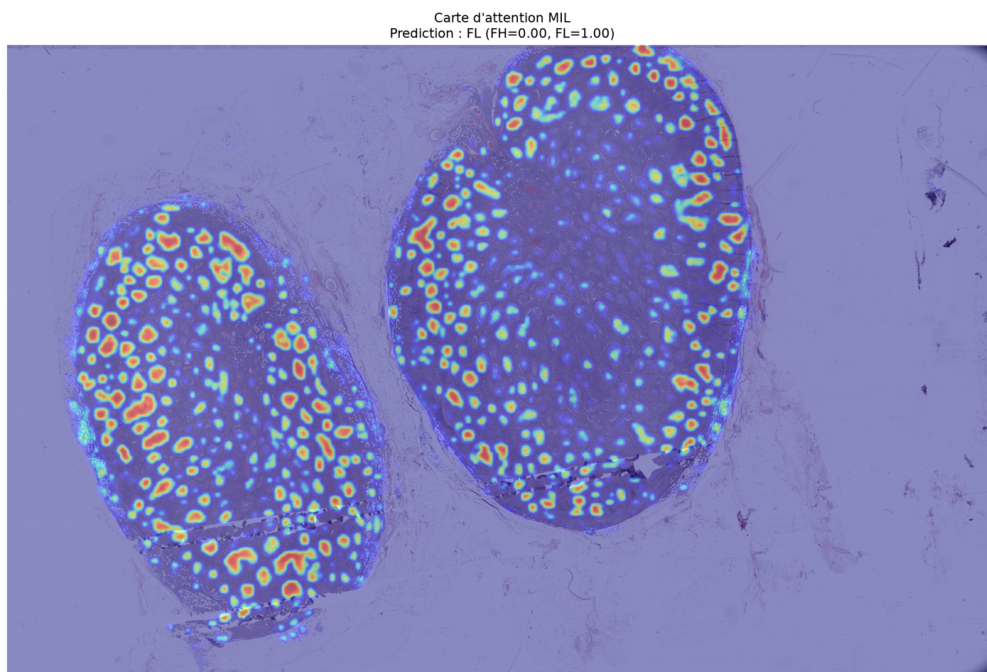


Rapport de stage — Aéro 4

Classification par intelligence artificielle d'hyperplasies et de lymphomes folliculaires

*Stage d'ingénieur en intelligence artificielle : développement et évaluation d'un modèle sur
lames virtuelles (Whole Slide Images) en anatomopathologie*



Simon LIGNAC — Aéro 4

Tuteurs en entreprise : Dr Camille FRANCHET
Stéphane TREILLARD

Tuteur pédagogique : Laetitia PAPION

Stage effectué du 3 août 2026
au 25 septembre 2026

Remerciements

Je souhaite tout d'abord remercier mon maître de stage, le Dr Camille Franchet, d'avoir accepté de m'accueillir comme stagiaire au sein de son équipe, au laboratoire d'anatomopathologie de l'Institut Universitaire du Cancer de Toulouse, et de m'avoir conseillé, aidé et accompagné tout au long de mon stage.

Je remercie ensuite Stéphane Treillard, doctorant en intelligence artificielle encadré par le Dr Franchet, qui m'a accompagné tout au long de mon stage et m'a expliqué les différentes approches que j'ai mises en œuvre, ainsi que Robin Schwob, également doctorant en intelligence artificielle dans l'équipe, qui, bien qu'il ne travaille pas sur le site de l'Oncopole, m'a consacré du temps pour m'aider.

Je tiens également à remercier le Dr Charlotte Syrykh, médecin spécialiste en anatomie et cytologie pathologiques, pour son aide et pour m'avoir expliqué son travail.

Sommaire

Remerciements	2
Fiche de synthèse	4
1 Introduction	5
2 L’IUCT-Oncopole, un centre où se rejoignent soin, enseignement et recherche contre le cancer	6
2.1 Présentation générale de l’établissement	6
2.2 Statut, gouvernance et instances représentatives	6
2.3 Le laboratoire d’anatomopathologie	7
2.4 Environnement de travail et encadrement	7
2.5 Stratégie de développement durable et RSE	8
3 Travail réalisé : un pipeline de classification par IA réutilisable	9
3.1 Problématique et démarche générale	9
3.2 Création et structure d’une lame numérisée	10
3.3 Environnement de travail	11
3.4 Le pipeline de traitement (commun aux deux jeux)	12
3.4.1 Découpage en patchs	12
3.4.2 Encodage par un modèle de fondation	13
3.4.3 Découpage des données par patient	13
3.4.4 Métriques d’évaluation : accuracy, AUC, perte	14
3.5 Phase 1 — Construction et validation de la méthode (jeu Hodgkin)	15
3.5.1 Socle commun : réseaux de neurones et entraînement	15
3.5.2 Modèle de référence : vote majoritaire	16
3.5.3 Comparaison des encodeurs : ResNet50 et H-optimus-0	16
3.5.4 Approche multi-échelle par combinaison de scores	17
3.5.5 Amélioration : le MIL avec attention	18
3.5.6 Résultats et interprétation	19
3.5.7 Exploration multi-échelle en concaténation ($20\times + 5\times$)	21
3.5.8 Améliorations de l’entraînement : early stopping et réduction adaptative du taux d’apprentissage	23
3.5.9 Cartes d’attention : visualiser où le modèle regarde	24
3.6 Phase 2 — Application au lymphome folliculaire	25
3.6.1 Approche adaptative par double encodeur	28
3.7 Pour aller plus loin : quelles dimensions portent l’information?	33
3.8 Discussion et comparaison avec la littérature	35
3.8.1 Comparaison avec les travaux de Syrykh <i>et al.</i> (2020)	35
3.8.2 Fusion multi-échelle par concaténation	36
3.8.3 Approche à double encodeur	36
3.9 Difficultés rencontrées	37
Conclusion et perspectives	38
Annexes	39
Bibliographie / Webographie	42
Liste des sigles et abréviations	44
Glossaire	45
Résumés	47

Fiche de synthèse

Sujet de stage	Classification automatique d'hyperplasie folliculaire (bénin) et de lymphome folliculaire (malin) par apprentissage profond sur lames virtuelles.
Objectifs	Construire un classifieur par intelligence artificielle permettant de distinguer l'hyperplasie folliculaire (bénin) du lymphome de Hodgkin (malin) dans un premier temps, puis réutiliser le pipeline pour détecter le lymphome folliculaire.
Démarche	Préparation des données (découpage en patches, encodage) puis entraînement des modèles sur les deux jeux de données (lymphome de Hodgkin, lymphome folliculaire).
Outils utilisés	Python, PyTorch, AtlasPatch (segmentation du tissu via SAM2 et encodage), modèles de fondation H-optimus-0, H-optimus-1 et Virchow2, ResNet50, scikit-learn, h5py, OpenSlide, Matplotlib; environnement WSL2/Ubuntu.
Études réalisées	Découpage des lames en patches aux grossissements 20× et 5×; encodage par modèle de fondation; classification par vote majoritaire; comparaison d'encodeurs (ResNet50 et H-optimus-0); Multiple Instance Learning (MIL); approches multi-échelles (moyenne des scores, puis concaténation 20× + 5×); cartes d'attention; double encodeur; sélection des features importantes.
Résultats (Hodgkin)	Forte progression sur le jeu Hodgkin (accuracy / AUC) : vote majoritaire ResNet50 0,746 / 0,866; vote majoritaire H-optimus-0 0,864 / 0,956; MIL mono-échelle H-optimus-1 0,932 / 0,977; jusqu'à 0,949 / 0,984 pour le MIL multi-échelle par concaténation.
Résultats (lymphome folliculaire)	Sur le jeu lymphome folliculaire (accuracy / AUC) : MIL mono-échelle H-optimus-1 0,946 / 0,992; MIL multi-échelle Virchow2 0,931 / 0,984; double encodeur (biopsie : H-optimus-1 20×, pièce chirurgicale : MIL multi-échelle Virchow2) 0,962 / 0,989, meilleur résultat avec 5 erreurs sur 130 lames et 2 lymphomes manqués sur 75.
Écarts / difficultés	Clé USB qui surchauffe; problèmes de stockage (environ 950 Go de données traitées pour 1 To de stockage); encodage très long (jusqu'à trois jours, soit environ 75 heures).
Reste à faire	Mettre en place la validation croisée et moyenner sur plusieurs seeds pour valider la règle du double encodeur; ajouter un modèle d'IA qui classe biopsie et pièce chirurgicale; utiliser la réduction de dimension (sélection de features) pour des modèles plus rapides.

1 Introduction

Le diagnostic du cancer repose en grande partie sur l'examen au microscope de tissus prélevés chez le patient, un travail qui est confié au pathologiste. Depuis peu, la numérisation de ces lames et les progrès de l'intelligence artificielle ouvrent une nouvelle voie : la pathologie computationnelle, qui consiste à appliquer l'apprentissage profond (deep learning) à l'analyse de lames numérisées afin d'aider au diagnostic. Ce domaine connaît une croissance rapide, causée par les avancées récentes de l'IA.

Mon stage s'inscrit dans ce domaine et porte sur la distinction entre deux diagnostics : l'hyperplasie folliculaire, une réaction bénigne du ganglion, et le lymphome folliculaire, un cancer qui se développe à partir des lymphocytes, des cellules du système immunitaire, qui touche fréquemment les ganglions lymphatiques et qui se décline en plusieurs grades. Les deux sont très similaires, ce qui rend leur distinction difficile, y compris pour un pathologiste. Développer un outil performant capable de les distinguer avec une grande précision est donc très utile pour un pathologiste et lui permettrait d'avoir une aide ou une confirmation à la décision du diagnostic.

J'ai réalisé ce stage à l'Institut universitaire du cancer de Toulouse (IUCT-Oncopole), au sein du laboratoire d'anatomopathologie, du 3 août au 25 septembre 2026, sous l'encadrement du Dr Camille Franchet et de Stéphane Treillard. J'y suis arrivé avec l'envie d'appliquer mes compétences d'ingénieur à un problème réel et porteur de sens, et de découvrir un domaine, celui de l'intelligence artificielle appliquée à la médecine, que je connaissais encore peu.

J'ai choisi ce stage car j'avais envie de découvrir de nouveaux domaines et, en discutant avec des médecins, je me suis rendu compte qu'il existait une réelle demande dans le domaine de l'intelligence artificielle appliquée à la médecine. Le sujet m'a tout de suite plu, car il me permettait d'appliquer mes compétences techniques en intelligence artificielle à un nouveau domaine, celui de la pathologie computationnelle.

Mon travail s'est déroulé en deux temps. J'ai d'abord construit et validé ma méthode sur un premier jeu de données, celui du lymphome de Hodgkin : un cancer du système lymphatique caractérisé par la présence de cellules tumorales particulières, dites cellules de Reed-Sternberg. Une fois la méthode éprouvée, je l'ai appliquée au jeu visé par le sujet du stage, le lymphome folliculaire. Ce passage d'un jeu de données à l'autre a été l'occasion de démontrer la robustesse et la réutilisabilité de mon approche.

Ce rapport s'organise en deux parties. La première présente l'établissement d'accueil, l'IUCT-Oncopole et son laboratoire d'anatomopathologie. La seconde décrit le travail réalisé : d'abord le pipeline de traitement commun aux deux jeux de données, puis les deux phases, à savoir la mise au point sur le lymphome de Hodgkin et l'application au lymphome folliculaire.

2 L'IUCT-Oncopole, un centre où se rejoignent soin, enseignement et recherche contre le cancer

2.1 Présentation générale de l'établissement

L'Institut Universitaire du Cancer de Toulouse-Oncopole (IUCT-Oncopole) est un établissement spécialisé dans la prise en charge des patients atteints de cancer. Ouvert en 2014 dans le sud de Toulouse, il réunit sur un même site environ 2 000 professionnels autour de trois spécialités : le soin, la recherche et l'enseignement. Il est labellisé Comprehensive Cancer Center, un label qui distingue les centres intégrant soins, recherche et formation en cancérologie.

L'IUCT-Oncopole est implanté sur le campus de l'Oncopole de Toulouse, construit sur l'ancien site de l'usine AZF, détruite par l'explosion de 2001. Ce campus rassemble des acteurs publics et privés de la lutte contre le cancer : on y trouve notamment le Centre de Recherches en Cancérologie de Toulouse (CRCT), qui regroupe des équipes de recherche fondamentale, ainsi que des partenaires industriels.

Un siècle de lutte contre le cancer à Toulouse. L'Institut Claudius Regaud, centre de lutte contre le cancer de Toulouse, a été fondé en 1923. Pendant près d'un siècle, il a fonctionné séparément des services de cancérologie du CHU de Toulouse. En 2014, les deux établissements ont regroupé leurs équipes sur un même site, l'Oncopole, pour créer l'IUCT-Oncopole.

L'activité en chiffres. En 2023, l'IUCT-Oncopole disposait de 308 lits, dont 103 places en hôpital de jour. Le site est occupé par environ 2 175 personnes, dont 191 médecins, 752 soignants, 147 chercheurs et 77 doctorants(IUCT-Oncopole, 2024).

Un positionnement plutôt qu'une concurrence. Il n'existe pas de réelle concurrence entre les hôpitaux publics et les CLCC. Les établissements ont pour objectif de couvrir les besoins de soins d'un territoire. L'IUCT-Oncopole est le centre de référence en cancérologie de la région Occitanie Ouest. Sur le plan national, il fait partie du réseau Unicancer, qui regroupe les 18 CLCC français. Entre ces centres, les relations sont surtout complémentaires : essais cliniques multicentriques, partage de données et recrutement de patients pour les maladies rares.

2.2 Statut, gouvernance et instances représentatives

L'IUCT-Oncopole fait partie d'un groupe de coopération sanitaire, qui comprend : l'Oncopole Claudius Regaud (OCR) et les services du CHU de Toulouse. L'OCR est un Centre de Lutte Contre le Cancer (CLCC), établissement de santé privé à but non lucratif fondé en 1923 et membre de la fédération Unicancer, qui regroupe les 18 CLCC français. La gouvernance du site est à l'image de ce groupe : le directeur général du CHU en assure la présidence et le directeur général de l'OCR la direction générale. Le personnel est également séparé en deux parties. Les salariés de l'OCR, qui relèvent du droit privé, disposent d'un Comité Social et Économique (CSE) propre, chargé du dialogue social, des questions de santé et de sécurité au travail, ainsi que des activités sociales et culturelles. Les agents du CHU relèvent quant à eux de la fonction publique hospitalière et des instances de dialogue social de leur établissement. Mon stage était pour ma part rattaché au CHU de Toulouse.

2.3 Le laboratoire d'anatomopathologie

Le laboratoire d'anatomie et cytologie pathologiques (ACP) analyse les prélèvements tissulaires et cellulaires issus des biopsies (prélèvements d'échantillons de petite taille) et des interventions chirurgicales en vue d'établir un diagnostic. En cancérologie, l'examen anatomopathologique confirme la nature tumorale d'une lésion, en précise le type histologique et le grade, et conditionne la prise en charge thérapeutique.

Le service assure l'ensemble de la chaîne, de la réception du prélèvement à l'interprétation diagnostique par le pathologiste, en passant par les étapes techniques de préparation des lames. Il dispose pour cela de plusieurs plateaux techniques :

- l'histologie, qui prépare et colore de fines coupes de tissu pour en révéler l'architecture au microscope ;
- l'immunohistochimie, qui utilise des anticorps pour détecter des protéines spécifiques dans les cellules et ainsi caractériser la lésion ;
- la biologie moléculaire, qui analyse l'ADN et l'ARN des cellules à la recherche d'anomalies génétiques.

Il dispose aussi d'une plateforme de numérisation des lames permettant l'analyse assistée par ordinateur, domaine dans lequel s'inscrit mon stage.

Le laboratoire en chiffres. Le département d'anatomie et cytologie pathologiques est dirigé par le Pr Pierre Brousset. Il est constitué des services de Purpan, de Rangueil et de l'Oncopole, et d'environ 30 pathologistes et 65 techniciens, ce qui en fait l'un des plus importants laboratoires d'anatomopathologie d'Europe. Il est membre, depuis 2010, du réseau national des centres de référence labellisé par l'INCa, et fait office de centre régional de référence pour quatre cancers rares, dont les lymphomes. Il coordonne notamment, avec Créteil, le réseau national Lymphopath, qui examine chaque année 12 000 cas de lymphomes.

Des projets de recherche en intelligence artificielle. Le laboratoire travaille sur l'IA depuis la fin des années 2010. Le projet APRIORICS a été créé en 2020, coordonné par mon maître de stage le Dr Camille Franchet, et vise à mieux caractériser les cancers du sein à partir de milliers de lames numérisées, en partenariat avec le Health Data Hub, Bpifrance et Thales. L'équipe a également publié en 2020 une étude sur la distinction entre lymphome folliculaire et hyperplasie folliculaire (Syrykh et al., 2020), dont mon stage reprend la problématique.

2.4 Environnement de travail et encadrement

J'ai effectué mon stage au sein du laboratoire d'anatomopathologie, sous la supervision du Dr Camille Franchet, médecin pathologiste, et de Stéphane Treillard, doctorant. La thèse de ce dernier porte sur la classification globale des lymphomes par intelligence artificielle, à partir des lames numérisées. Mon travail s'inscrit comme une sous-partie de ce projet plus large : je me suis concentré sur un sous-problème de sa thèse : la distinction entre hyperplasie folliculaire et lymphome folliculaire.

Cette double supervision m'a permis de bénéficier de deux regards complémentaires. Le Dr Franchet m'a apporté l'expertise médicale et les difficultés que ces cas posent. Stéphane Treillard m'a aidé sur la méthodologie, l'explication des techniques d'apprentissage profond et sur la programmation.

Mon stage s'est déroulé du lundi au vendredi de 9h à 17h. Une réunion tous les jeudis à 17h avec Stéphane Treillard et Camille Franchet me permettait de présenter mes avancées et de discuter de mes résultats, tout en fixant les objectifs pour les semaines suivantes.

2.5 Stratégie de développement durable et RSE

Le CHU a engagé un premier diagnostic environnemental dès 2005 et s'est doté d'une direction du développement durable en 2010. Sa politique RSE, inscrite dans son projet d'établissement, s'appuie sur les 17 objectifs de développement durable de l'ONU et aborde les achats écoresponsables, les économies d'énergie, l'alimentation durable, la mobilité et la gestion des déchets ; elle fait l'objet d'un rapport annuel (CHU de Toulouse, 2024).

Un plan de mobilité propose aux agents des alternatives à la voiture individuelle pour les trajets domicile-travail et les déplacements professionnels (transports en commun, vélo, covoiturage). Des projets d'éco-conception des soins, comme « Green bloc » dans les blocs opératoires et « Green réa » en réanimation, visent à réduire la production de déchets et à mieux les valoriser. Cette démarche a été récompensée en avril 2026 : le CHU de Toulouse est devenu le premier établissement français à obtenir le label THQSE (Très Haute Qualité Sanitaire, Sociale et Environnementale) « chirurgie responsable » au niveau Or, avec un score de 92 %. Sur le plan social, la RSE passe également par le dialogue social porté par les instances représentatives décrites plus haut.

L'impact environnemental de mon propre travail L'impact environnemental de mon travail reste relativement limité. Il provient essentiellement de la consommation énergétique liée à l'utilisation de l'ordinateur et au lancement de programmes pouvant durer jusqu'à trois jours sans interruption, consommant de l'électricité.

3 Travail réalisé : un pipeline de classification par IA réutilisable

3.1 Problématique et démarche générale

Mon stage porte sur l'utilisation de l'apprentissage profond pour analyser des lames virtuelles. L'apprentissage profond (deep learning) est un sous-domaine de l'apprentissage automatique qui s'appuie sur des réseaux de neurones composés de nombreuses couches successives. Sa spécificité est d'apprendre directement à partir des données brutes, ici les pixels d'une lame numérisée, les caractéristiques pertinentes pour la tâche, sans qu'un expert ait à les définir manuellement.

Le problème traité est de distinguer une hyperplasie folliculaire (FH, de l'anglais follicular hyperplasia), réaction bénigne du ganglion, d'un lymphome folliculaire (FL, follicular lymphoma), qui est un cancer. Comme le montre la figure 3.1, les deux présentent des follicules sur les lames colorées en H&E, la coloration standard réalisée systématiquement, et leur différenciation reste difficile, en particulier pour les lymphomes de bas grade. Pour trancher, le pathologiste s'appuie sur des marquages complémentaires en immunohistochimie, comme BCL2 : les follicules restent clairs dans l'hyperplasie, alors que les cellules tumorales du lymphome sont marquées en brun. L'objectif de mon travail est donc de déterminer si un modèle d'IA peut faire cette distinction à partir de la seule lame H&E, en prenant une lame entière en entrée pour donner un diagnostic FH ou FL, et ainsi fournir au pathologiste un deuxième avis avec un score de confiance dès cette première étape.

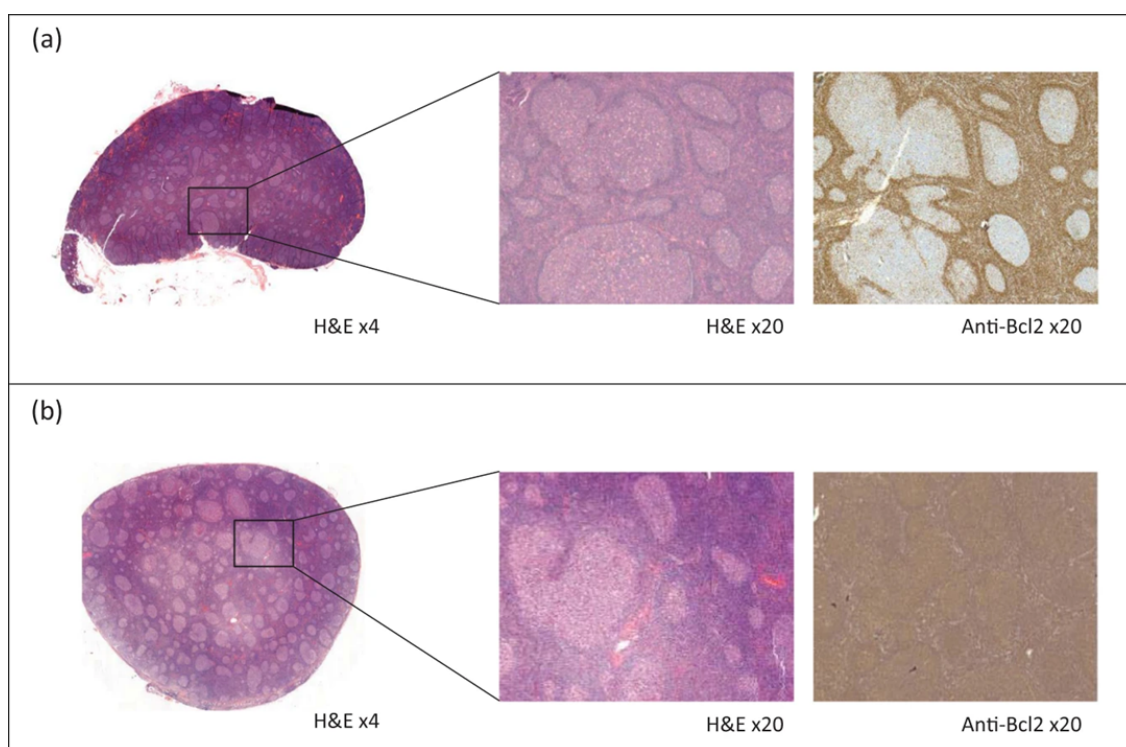


FIGURE 3.1 – Distinction histopathologique entre hyperplasie folliculaire (FH) (a) et lymphome folliculaire (FL) (b) : coloration H&E aux grossissements 4 $\times$ et 20 $\times$, puis marquage anti-BCL2 à 20 $\times$. Source : Syrykh et al., *npj Digit. Med.*, 2020, sous licence CC BY 4.0.

Le stage s'est déroulé en deux temps, avec deux jeux de données. J'ai d'abord construit mon pipeline et entraîné mon modèle sur un autre type de lymphome, le lymphome de Hodgkin, plus facilement détectable, puis je l'ai appliqué au lymphome folliculaire, ce qui a permis de montrer sa réutilisabilité.

3.2 Création et structure d'une lame numérisée

Dans un premier temps, j'ai bénéficié d'une visite du laboratoire et d'une explication de toutes les étapes, du passage du prélèvement en bloc opératoire jusqu'à la lame. Ce procédé est réparti en plusieurs étapes :

Réception / enregistrement Le prélèvement arrive d'un sas depuis le bloc opératoire jusqu'au laboratoire. Il est enregistré et reçoit un numéro d'identification unique.

Macroscopie La pièce est ensuite observée, puis découpée en petits prélèvements placés dans des cassettes.

Fixation Les tissus sont fixés dans du formol pour stopper l'autolyse (la dégradation des cellules après la mort du tissu) et figer les structures cellulaires.

Inclusion en paraffine Les cassettes sont déshydratées dans des bains d'éthanol puis mises dans de la paraffine qui se solidifie pour être ensuite coupée finement.

Coupe (microtomie) Le bloc est coupé en lamelles très fines (3 à 5 μm). Celles-ci sont ensuite posées sur des lames de verre. À ce stade, le tissu est incolore donc invisible au microscope.

Coloration et montage On applique ensuite des colorants sur les lames. La coloration standard utilisée en anatomopathologie est la coloration H&E (hématoxyline-éosine). Celle-ci utilise deux colorants : l'hématoxyline qui colore les noyaux en bleu et violet, et l'éosine, qui colore les cytoplasmes et le tissu conjonctif en rose. Une autre lamelle de verre est placée par-dessus pour refermer la lame. On obtient en résultat une lame lisible au microscope.

Morphologie (lecture diagnostique) Le pathologiste regarde la lame au microscope, observe l'architecture à différentes échelles et la cytologie, c'est-à-dire l'aspect des cellules elles-mêmes (forme, taille, noyau), puis pose son diagnostic. Dans le cadre de mon stage, c'est à ce moment-là que la distinction entre hyperplasie folliculaire et lymphome folliculaire est faite.

Immunohistochimie (et microscopie électronique) Lorsque la morphologie ne permet pas de trancher, on réalise des marquages immunohistochimiques : des anticorps révèlent des protéines spécifiques et précisent le diagnostic. Par exemple, les lymphocytes B et T sont impossibles à distinguer sur une simple lame H&E, car ils ont le même aspect au microscope. On les différencie grâce à des marquages ciblant des protéines caractéristiques de chaque type : le CD3, présent à la surface des lymphocytes T, et le CD20, présent sur les lymphocytes B. D'autres marqueurs (BCL2, CD10, BCL6...) sont utilisés pour caractériser les lymphomes.

Biologie moléculaire Sur le tissu (souvent en FFPE, fixé-inclus en paraffine), on recherche des anomalies moléculaires : mutations, réarrangements, translocations.

Numérisation des lames (plateforme d'imagerie) Les lames colorées sont scannées à haute résolution par un scanner de lames, ce qui produit des images numériques (WSI, whole slide images) exploitables à plusieurs grossissements (5 $\times$, 20 $\times$, 40 $\times$...). Ce sont précisément ces images que je vais traiter au cours de mon stage.

Une lame numérisée (WSI) n'est pas stockée comme une seule image géante, mais sous forme de pyramide multi-résolution, comme l'illustre la figure 3.2. La même lame est enregistrée à plusieurs niveaux de grossissement : à la base de la pyramide, l'image au plus fort grossissement (40 $\times$), très détaillée mais très lourde ; au sommet, une version fortement réduite (1 $\times$), légère et grossière, qui

sert de vue d'ensemble; entre les deux, des niveaux intermédiaires (par exemple $10\times$). Chaque niveau est de plus découpé en petites tuiles.

Le principe est le même que celui de Google Earth : quand on regarde la Terre entière, l'application n'affiche qu'une image basse résolution du globe; à mesure que l'on zoome sur une ville puis sur une rue, elle ne charge que les tuiles haute résolution de la zone regardée, et non la planète entière en pleine définition. Une lame fonctionne exactement ainsi : selon le grossissement demandé et la région visée, seuls le niveau de la pyramide et les tuiles utiles sont chargés, ce qui permet de naviguer fluidement dans une image de plusieurs gigaoctets sans jamais la charger intégralement en mémoire.

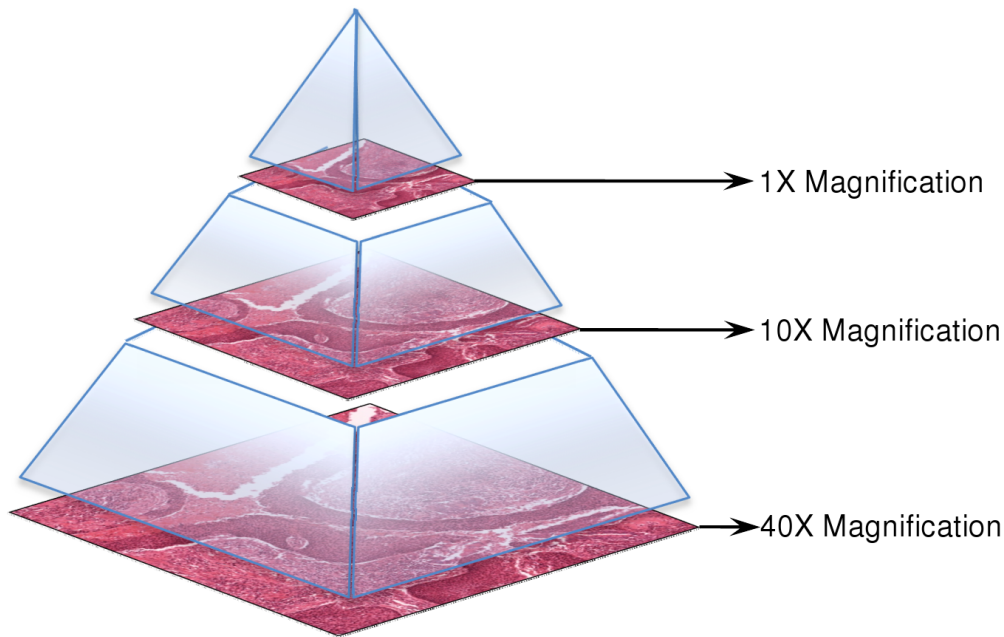


FIGURE 3.2 – Structure pyramidale de stockage d'une lame numérisée (WSI). Le niveau de grossissement le plus élevé ($40\times$) est conservé, ainsi que des niveaux de plus faible résolution ($10\times$ et $1\times$). Source : Wang et al. (2012), figure 5, sous licence CC BY.

3.3 Environnement de travail

Le laboratoire m'a confié un ordinateur portable avec la puissance nécessaire pour entraîner mes modèles (RTX 5090 Laptop GPU (24 Go de VRAM), Intel Core Ultra 9 275HX et 64 Go de RAM, sous WSL2/Ubuntu avec un environnement conda dédié et PyTorch) ainsi qu'une clé USB contenant 591 lames numérisées. Ce nombre peut paraître faible pour entraîner un modèle, mais il est à relativiser au vu de la taille de chaque lame. En effet, chaque lame possède une résolution très élevée du fait de son architecture pyramidale et, par conséquent, une taille très importante (une seule lame pouvait atteindre 4 Go). Les 591 lames représentaient ainsi un peu plus de 900 Go de stockage sur la clé USB. La clé USB contenait également un fichier Excel avec les noms de chaque lame, leur diagnostic et le type de prélèvement (biopsie ou pièce chirurgicale).

3.4 Le pipeline de traitement (commun aux deux jeux)

Cette section décrit l'ensemble des méthodes utilisées pour le découpage en patches, l'encodage, le découpage des données et l'évaluation par lame. Le pipeline est identique pour le lymphome de Hodgkin et pour le lymphome folliculaire. Seules les données changent, ce qui permet de valider la méthode sur le premier jeu puis de l'appliquer au second.

Deux jeux de données successifs

Deux jeux de données se sont succédé au cours du stage, dans un ordre chronologique qui reflète la démarche adoptée. Le premier, constitué de lames de lymphome de Hodgkin, a servi de support à la mise au point et à la validation de la méthode (Phase 1). Le second, constitué de lames de lymphome folliculaire, est la cible finale à laquelle la méthode est appliquée.

Jeu 1 Lymphome de Hodgkin (mise au point) Ce premier jeu comprend 591 lames numérisées de 591 patients différents (le découpage par patient décrit en section 3.4.3 n'a donc pas d'effet sur ces données, mais il a été appliqué par rigueur et pour garantir la réutilisabilité du pipeline). Les lames se répartissent entre 280 hyperplasies folliculaires (FH) et 311 lymphomes de Hodgkin. Le protocole de découpage par patient les divise en 355 lames d'entraînement, 118 de validation et 118 de test, l'ensemble de test se composant de 51 hyperplasies et 67 lymphomes. Le découpage en patches produit au total 5 645 588 patches au grossissement $20\times$ (3 206 492 en entraînement, 1 221 443 en validation et 1 217 653 en test) et 374 946 au grossissement $5\times$ (respectivement 213 370, 80 839 et 80 737). C'est sur ce jeu qu'a été développée puis validée l'intégralité de la méthode décrite dans ce chapitre (résultats de la Phase 1).

Jeu 2 Lymphome folliculaire (cible) Le second jeu, fourni ultérieurement, comprend 651 lames opposant l'hyperplasie folliculaire (FH, 279 lames) au lymphome folliculaire (FL, 372 lames). Le même protocole de découpage par patient le partitionne en 391 lames d'entraînement, 130 de validation et 130 de test, l'ensemble de test comptant 55 hyperplasies et 75 lymphomes. Le découpage en patches produit au total 6 917 343 patches au grossissement $20\times$ (4 058 863 en entraînement, 1 428 014 en validation et 1 430 466 en test) et 459 044 au grossissement $5\times$ (respectivement 269 493, 94 477 et 95 074).

3.4.1 Découpage en patches

La première étape de mon stage a été de découper les lames en patches, car les lames brutes sont trop grandes pour être traitées par un modèle de machine learning. Dans un premier temps, j'ai choisi comme niveau de résolution le $20\times$ car il constitue un bon compromis : l'image n'est pas trop zoomée, tout en conservant une information suffisamment précise à l'échelle du patch. C'est aussi le niveau sur lequel sont entraînés la plupart des modèles de fondation (des modèles pré-entraînés sur d'énormes quantités d'images, réutilisables comme base pour de nombreuses tâches). J'ai ensuite lancé mon code avec une taille de patch de 256 par 256 pixels. J'ai choisi d'utiliser AtlasPatch (Alagha et al., 2026) pour mon stage, car il segmente le tissu en une seule fois sur une vignette via SAM2 (Ravi et al., 2024), un modèle de fondation de Meta dédié à la segmentation d'images, parallélise le traitement et ne stocke que les coordonnées des patches utiles (avec du tissu dessus). Cela le rend à la fois beaucoup plus rapide et plus fidèle dans la détection du tissu, ce qui en fait l'un des outils de préprocessing les plus performants aujourd'hui. Pour des raisons de mémoire et de rapidité, le nombre de patches par lame est plafonné de 500 à 2000 en fonction de l'architecture utilisée (500 en vote majoritaire, 2000 en MIL).

3.4.2 Encodage par un modèle de fondation

AtlasPatch se charge aussi de l’encodage, qui consiste à utiliser un modèle de fondation (un réseau de neurones entraîné en auto-supervision sur de très grandes quantités d’images, qui produit pour chaque patch un vecteur de features réutilisable pour la classification) pour lire chaque patch et sortir un vecteur de features qui le décrit. Ces vecteurs sont ensuite prêts à être injectés et traités dans un réseau de neurones dont le rôle est de les agréger pour prédire un diagnostic à l’échelle de la lame entière. J’ai d’abord utilisé ResNet50 (He et al., 2016), un réseau de neurones convolutif (CNN). Cette famille d’architectures est plus ancienne, et surtout ce modèle a été pré-entraîné sur ImageNet (Deng et al., 2009), un ensemble d’images naturelles (animaux, objets, scènes du quotidien). ResNet50 constituait donc une première approche simple, mais peu spécialisée pour la pathologie. J’ai ensuite employé H-optimus-0 (Bioptimus, 2024), puis son successeur H-optimus-1 (Bioptimus, 2025), ainsi que Virchow2 (Zimmermann et al., 2024), des modèles de fondation dédiés à la pathologie. Ce sont des transformers (Vision Transformers ; Dosovitskiy et al., 2021), une architecture plus récente et très performante pour ce type de tâche, pré-entraînés sur d’immenses ensembles de lames numérisées. Cette spécialisation leur permet d’extraire des features bien plus pertinentes que celles d’un CNN généraliste. Chaque patch est alors représenté par un vecteur de features, prêt à être utilisé dans l’étape de classification.

3.4.3 Découpage des données par patient

La constitution des ensembles d’entraînement, de validation et de test obéit à une contrainte nécessaire : éviter toute fuite de données entre patients. En effet, dans un jeu de données, plusieurs lames peuvent appartenir à un même patient, et chaque lame peut présenter des caractéristiques propres à l’individu (particularités morphologiques, teintes...). Ce découpage est très important car si une lame d’un même patient se retrouvait dans le dataset d’entraînement et une autre lame de ce même patient dans le test, le modèle pourrait apprendre à reconnaître le patient plutôt que le diagnostic et obtiendrait donc des résultats très élevés, mais qui ne reposent pas sur un diagnostic réel du modèle. Cette situation peut se comparer à un élève qui apprendrait par cœur les réponses de plusieurs QCM : si le jour de l’examen, les questions sont celles d’un QCM de l’année précédente, il obtiendra une excellente note, sans pour autant maîtriser le fond du sujet. Si le QCM était recréé à partir de rien, il obtiendrait une mauvaise note. C’est pour cela que le découpage est réalisé au niveau du patient, et jamais au niveau de la lame ni du patch.

La séparation des données en entraînement, validation et test s’appuie sur `StratifiedGroupKFold` de scikit-learn, qui combine deux garanties simultanées.

D’une part, `groups=case_id` impose que toutes les lames d’un même patient restent regroupées dans un unique pli. D’autre part, `label` préserve les proportions des deux classes (FH et FL) dans chaque pli, ce qui permet au modèle de s’entraîner sur une proportion correcte d’hyperplasies et de lymphomes. Le jeu est donc divisé en cinq plis.

À partir de ces cinq plis, la répartition finale est obtenue en assignant le pli de test du premier découpage à l’ensemble de test (environ vingt pour cent), le deuxième à l’ensemble de validation (environ vingt pour cent), et les trois derniers à l’entraînement (environ soixante pour cent). Comme les plis de test produits par `StratifiedGroupKFold` sont disjoints, les ensembles de test et de validation ne se recouvrent pas, ce qui garantit un découpage d’environ 60/20/20.

La reproductibilité du protocole est assurée par une seed fixée : le tirage des plis est déterministe et les plis sont identiques à chaque exécution. Cette propriété est indispensable pour que les expériences soient comparables entre elles. Deux manifestes sont alors produits : ce sont des fichiers récapitulatifs listant toutes les données et leurs métadonnées (identifiant, patient, étiquette, ensemble d’appartenance). Un manifeste est produit au niveau des lames, utilisé pour piloter l’entraînement, et un autre au niveau des patches.

3.4.4 Métriques d'évaluation : accuracy, AUC, perte

L'accuracy représente la proportion de lames bien classées par le modèle.

L'AUC pour Area Under the Curve mesure la capacité de discrimination du modèle, indépendamment de tout seuil. Celle-ci correspond à la probabilité de donner un meilleur score à un lymphome folliculaire qu'à une hyperplasie. Une valeur de 0,5 représente une distinction aléatoire, tandis qu'un score de 1 est une séparation parfaite des deux classes.

La perte (loss) joue un rôle assez différent des métriques précédentes : c'est la grandeur que le modèle cherche à minimiser durant l'entraînement, et une valeur plus faible reflète donc une meilleure performance. La perte utilisée est l'entropie croisée, appliquée aux deux logits de sortie via un softmax – une formulation équivalente à l'entropie croisée binaire dans le cas à deux classes. Pour une lame de vraie classe c , dont le modèle prédit la distribution de probabilités (p_k) , elle s'écrit :

$$\mathcal{L} = - \sum_k y_k \log(p_k) = - \log(p_c), \quad (3.1)$$

la seconde égalité découlant du fait que l'étiquette y ne vaut 1 que pour la vraie classe c . Contrairement à l'accuracy, la perte prend en compte la confiance du modèle : une prédiction correcte mais peu assurée n'est pas beaucoup pénalisée, tandis qu'une prédiction fautive et très confiante l'est fortement. La perte constitue donc une mesure de la qualité des prédictions du modèle. Cette différence de nature explique un phénomène observé en pratique : la perte de validation peut augmenter alors que l'accuracy et l'AUC restent constantes ou progressent. Le modèle devient en effet de plus en plus confiant dans ses erreurs, créant un surapprentissage de calibration que seule la perte permet de détecter.

Les trois grandeurs précédentes servent surtout à suivre l'entraînement au fil des epochs, une epoch désignant un passage complet du modèle sur l'ensemble des données d'entraînement. L'évaluation finale du modèle retenu s'appuie en complément sur quatre métriques calculées par classe sur le jeu de test.

La précision mesure, parmi les lames prédites dans une classe, la proportion de bonnes prédictions ; elle sanctionne les faux positifs. Le rappel représente la proportion de lames bien prédites parmi les lames appartenant réellement à une classe. Il sanctionne donc les faux négatifs. Le F1-score est une moyenne harmonique des deux et le support est le nombre de lames dans chaque classe. Enfin la moyenne macro considère que les deux classes sont équitables tandis que la moyenne pondérée prend en compte le support des deux classes.

3.5 Phase 1 — Construction et validation de la méthode (jeu Hodgkin)

La phase 1 aborde la conception et la validation des différents modèles sur le jeu Hodgkin. C'est sur ce jeu que ma méthode a été codée et testée en premier. Il faut toutefois souligner que les résultats présentés dans cette section valident la méthode sur le problème « hyperplasie folliculaire contre lymphome de Hodgkin ». Le lymphome de Hodgkin, plus facilement détectable, a ici servi de support pour s'assurer que l'approche était solide avant de l'appliquer au jeu cible. Dans cette phase, la classe maligne est notée « Hodgkin ».

3.5.1 Socle commun : réseaux de neurones et entraînement

Tous les modèles que j'ai développés au cours de mon stage reposent sur un réseau de neurones. Un réseau de neurones est constitué d'un empilement de couches qui transforment un vecteur d'entrée (ici les features d'un ou plusieurs patches) en un vecteur de sortie (ici deux scores, un par classe).

Les couches linéaires constituent la brique élémentaire du réseau. Une telle couche transforme le vecteur x qu'elle reçoit en un vecteur de dimension différente selon l'opération :

$$y = Wx + b, \quad (3.2)$$

où la matrice de poids W et le vecteur de biais b sont précisément les paramètres que le réseau ajuste au cours de l'entraînement. Entre les couches, on rajoute deux éléments indispensables. Le premier est la fonction d'activation ReLU (*Rectified Linear Unit*), qui ne conserve que les valeurs positives :

$$\text{ReLU}(x) = \max(0, x). \quad (3.3)$$

Sans une telle fonction, empiler des couches linéaires reviendrait à une seule transformation linéaire, incapable de séparer des classes qui ne le sont pas ; en introduisant une non-linéarité, la ReLU permet au réseau d'apprendre des frontières de décision complexes. Le second élément est le Dropout, qui met à zéro aléatoirement une partie des neurones (ici 30 %) durant l'entraînement, ce qui permet au réseau de ne pas se baser seulement sur quelques neurones, limitant ainsi le surapprentissage.

Ces briques de base sont communes à tous mes modèles ; seul leur agencement change de l'un à l'autre.

Le mécanisme d'apprentissage, lui aussi, est commun à tous : la descente de gradient. À chaque étape, le modèle traite un lot de données et produit ses prédictions, puis une fonction de perte (ici l'entropie croisée) mesure l'écart entre ces prédictions et les vrais labels. On calcule alors le gradient de cette perte par rapport à chaque poids, $\nabla_W \mathcal{L}$: il indique dans quelle direction, et avec quelle ampleur, modifier chaque poids pour diminuer l'erreur. Chaque poids est ensuite corrigé d'un petit pas dans la direction opposée au gradient :

$$W \leftarrow W - \eta \nabla_W \mathcal{L}, \quad (3.4)$$

où le taux d'apprentissage η (learning rate) fixe la taille du pas. En répétant cette opération sur tous les lots, epoch après epoch, le réseau converge progressivement vers des poids qui minimisent l'erreur.

Algorithm 1 : Une étape de descente de gradient

Input : Un lot (x_b, y_b) , le modèle, la fonction de perte, l'optimiseur
 Réinitialiser les gradients accumulés ;
 Calculer les prédictions du modèle sur x_b ;
 Évaluer la perte $\mathcal{L}$ entre ces prédictions et les vrais labels y_b ;
 Rétropropager $\mathcal{L}$ pour obtenir les gradients $\nabla_W \mathcal{L}$;
 Mettre à jour chaque poids : $W \leftarrow W - \eta \nabla_W \mathcal{L}$;

J'utilise l'optimiseur AdamW plutôt qu'une descente de gradient classique : il adapte automatiquement le pas d'apprentissage à chaque poids et intègre une régularisation (weight decay). La régularisation désigne l'ensemble des techniques qui contraignent le modèle durant l'entraînement pour l'empêcher de trop coller aux données vues, et améliorer ainsi sa capacité à généraliser à de nouvelles données. Ici, le weight decay pénalise les poids de trop grande amplitude, ce qui limite le surapprentissage.

3.5.2 Modèle de référence : vote majoritaire

Pour le vote majoritaire, j'utilise un perceptron multicouche (MLP) qui classe chaque patch individuellement : il prend en entrée le vecteur de features d'un patch et produit en sortie deux scores, un par classe (bénin et malin). Il est constitué de trois couches linéaires successives qui réduisent progressivement la dimension du vecteur : de 2048 pour ResNet50 vers 512, puis 512 vers 128, et enfin 128 vers le nombre de classes. On y retrouve les briques décrites précédemment : une activation ReLU après chaque couche et un Dropout de 0,3 après chacune des deux premières couches.

Le vote majoritaire consiste à déterminer le diagnostic de la lame à partir du niveau des patches : pour une lame donnée, on compte le nombre de patches considérés comme malins (c'est-à-dire avec un score supérieur à 0,5) et le nombre de patches bénins. On compare ensuite les deux nombres ; si les patches malins sont majoritaires, la lame est considérée comme maligne, et inversement dans le cas contraire.

Ce modèle est simple et efficace, mais il présente un inconvénient majeur : il ne prend pas en compte la répartition spatiale de la tumeur. Certains lymphomes, par exemple, sont très concentrés en quelques zones précises sans s'étendre à l'ensemble de la lame ; ce type de cas risque alors de ne pas être détecté par le vote majoritaire.

3.5.3 Comparaison des encodeurs : ResNet50 et H-optimus-0

J'ai comparé deux encodeurs pour la même approche par vote majoritaire, à grossissement identique (20×) et sur le même jeu de test (118 lames du jeu Hodgkin). Le premier, ResNet50, est un réseau de neurones convolutif pré-entraîné sur ImageNet, une base d'images naturelles généralistes. Le second, H-optimus-0, est un modèle de fondation de type transformer, pré-entraîné spécifiquement sur de grandes quantités d'images histologiques.

TABLE 3.1 – Comparaison des encodeurs (vote majoritaire, 20×, test), jeu Hodgkin.

Encodeur	Accuracy	AUC
ResNet50 (ImageNet)	0,746	0,866
H-optimus-0	0,864	0,956

Le passage de ResNet50 à H-optimus-0 fait progresser l'accuracy de 0,746 à 0,864 et l'AUC de 0,866 à 0,956. Cette comparaison permet de montrer l'importance de l'utilisation d'un encodeur pré-entraîné sur des images de pathologie numérisées. H-optimus-0 produit des features plus adaptées à la détection de la différence entre tissu bénin et tissu tumoral qu'un réseau généraliste entraîné sur des images naturelles. Les courbes d'apprentissage (figures 3.14 et 3.15) et les matrices de confusion (figure 3.3) confirment ce constat : ResNet50 commet davantage d'erreurs, alors que H-optimus-0 est plus performant dans les deux classes. Pour la suite, j'ai donc retenu H-optimus-0 comme encodeur.

Les courbes d'apprentissage complètes de H-optimus-0 et ResNet50 sont reportées en annexe (figures 3.14 et 3.15).

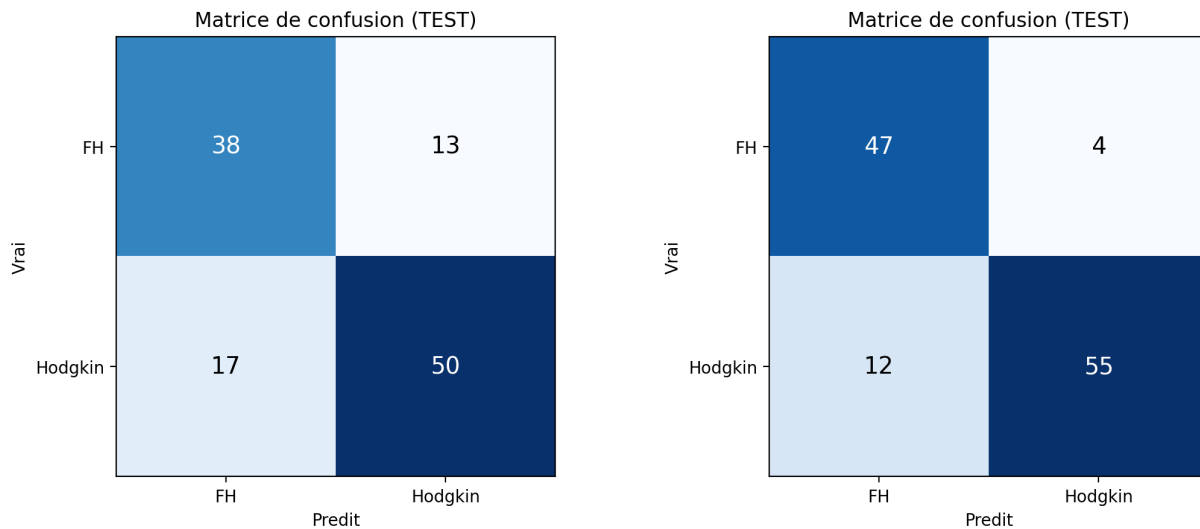


FIGURE 3.3 – Matrices de confusion sur le test : ResNet50 (à gauche) et H-optimus-0 (à droite), vote majoritaire à $20\times$, jeu Hodgkin.

3.5.4 Approche multi-échelle par combinaison de scores

J’ai été amené à chercher une piste d’amélioration au cours de mon stage. En effet, lorsqu’un pathologiste regarde une lame au microscope, il change constamment de grossissement. La lame est examinée à plusieurs échelles, car chacune apporte une information différente. Le faible grossissement ($5\times$) donne une vue de l’architecture, tandis que le fort grossissement ($20\times$) révèle plus de détails. Cela m’a orienté vers l’idée de reproduire ce principe dans mon modèle, en lui faisant exploiter cette nouvelle dimension d’échelle.

Ma première idée pour intégrer cette information multi-échelle a été de combiner les prédictions obtenues aux deux grossissements. J’entraîne deux modèles indépendants, identiques dans leur architecture : l’un sur les patchs extraits à $20\times$, l’autre sur les patchs extraits à $5\times$. Chaque modèle produit, pour une lame donnée, un score correspondant à la probabilité moyenne qu’elle soit maligne sur l’ensemble de ses patchs. Le score final de la lame est alors la moyenne des deux scores. Cette combinaison de plusieurs modèles est une approche répandue en apprentissage automatique et en pathologie computationnelle (Campanella et al., 2019).

TABLE 3.2 – Combinaison des échelles (vote majoritaire, H-optimus-0, test), jeu Hodgkin. Les deux modèles ont été réentraînés pour la fusion (8 epochs chacun), d’où des valeurs légèrement différentes de celles des tableaux 3.1 et 3.4.

Échelle	Accuracy	AUC
$20\times$ seul	0,873	0,947
$5\times$ seul	0,847	0,935
Moyenne $5\times + 20\times$	0,864	0,946

Le résultat est décevant mais assez prévisible : la moyenne des deux échelles (0,864) ne dépasse pas le $20\times$ utilisé seul (0,873), elle lui est même légèrement inférieure. La moyenne des scores n’apporte donc aucune plus-value. Cela est dû au fait que le $5\times$ (0,847) est moins performant que le $20\times$ (0,873) et le tire vers le bas. L’explication tient à l’indépendance de l’entraînement des deux modèles. En effet, le $5\times$ a beaucoup plus de mal à voir l’information car, bien qu’il voie l’architecture de la lame, il ne peut pas explorer les détails : il n’a qu’une vue assez grossière de la lame. C’est pourquoi, dans la suite de mon stage, j’utilise une méthode qui combine les deux informations en

entraînant les deux branches ensemble.

Les courbes d'apprentissage des deux échelles (5× et 20×) sont reportées en annexe (figure 3.16).

Algorithm 2 : Combinaison des deux échelles : moyenne des scores (fusion tardive)

Input : Les manifestes des deux échelles, $\mathcal{M}_{20\times}$ et $\mathcal{M}_{5\times}$
Output : Un score de fusion par lame de test
 Entraîner un modèle sur les patches 20×, puis prédire les scores de test $s^{20\times}$;
 Entraîner un modèle sur les patches 5×, puis prédire les scores de test $s^{5\times}$;
foreach lame ℓ du jeu de test **do**
 Apparier les scores des deux échelles pour la lame ℓ ;
 $s_{\ell}^{\text{fusion}} \leftarrow \frac{s_{\ell}^{20\times} + s_{\ell}^{5\times}}{2}$;
return les scores de fusion s^{fusion}

3.5.5 Amélioration : le MIL avec attention

Pour dépasser les limites du vote majoritaire, je me suis tourné vers une approche appelée MIL (Multiple Instance Learning, ou apprentissage à instances multiples). Celui-ci consiste à traiter chaque lame non plus comme un ensemble de patches jugés indépendamment, mais comme un tout : le modèle reçoit l'ensemble des patches d'une lame et apprend lui-même à produire une décision unique au niveau de la lame.

La plus grande différence avec le vote majoritaire vient du traitement des patches. Avec le vote majoritaire, chaque patch possède le même poids que les autres : un patch porteur d'information compte autant dans la décision qu'un patch qui n'en porte pas. Le souci est qu'une tumeur concentrée dans quelques patches peut alors se retrouver mal diagnostiquée. Le MIL corrige ce défaut grâce à un mécanisme d'attention : plutôt que de compter les patches, le modèle apprend à attribuer à chacun un poids d'attention en fonction de son importance (Ilse et al., 2018). L'attention désigne un mécanisme par lequel un réseau apprend à pondérer les éléments qu'il reçoit, afin de concentrer son traitement sur les plus pertinents pour la tâche et de négliger les autres. Une petite zone tumorale peut ainsi recevoir un poids fort et suffire à faire basculer la décision.

Chaque patch est d'abord transformé par un encodeur en une représentation interne, à partir de laquelle une branche d'attention calcule un score d'importance. Ces scores sont normalisés par une fonction softmax, qui les transforme en poids tous positifs et de somme égale à 1 (à la manière d'une distribution de probabilités). La représentation de la lame est alors obtenue comme la moyenne des représentations des patches pondérée par ces poids. C'est ce vecteur unique, qui résume la lame en insistant sur ses régions les plus informatives, qui est ensuite classé.

Algorithm 3 : Branche d'attention (encodage, attention, agrégation)

Input : Les patches d'une lame $P = (p_1, \dots, p_N)$
Output : Le vecteur agrégé de la lame z et les poids d'attention
foreach patch p_i **do**
 $h_i \leftarrow \text{Encodeur}(p_i)$; // projection 1536 → 256, ReLU, Dropout
 $s_i \leftarrow \text{Attention}(h_i)$; // score d'importance scalaire
 $(a_1, \dots, a_N) \leftarrow \text{softmax}(s_1, \dots, s_N)$; // poids positifs, de somme 1
 $z \leftarrow \sum_{i=1}^N a_i h_i$; // moyenne des patches pondérée par l'attention
return $z, (a_i)$

L'agrégation pondérée des patches s'écrit :

$$z = \sum_{i=1}^N a_i h_i, \quad a_i = \frac{\exp(s_i)}{\sum_{j=1}^N \exp(s_j)}, \quad (3.5)$$

où h_i est la représentation du patch i et z le vecteur final de la lame. Le score d'attention s_i est une valeur réelle produite par la branche d'attention à partir de la représentation h_i : il traduit l'importance accordée au patch i pour la décision, un score élevé signalant un patch jugé informatif. Ces scores sont ensuite normalisés par le softmax en poids a_i , positifs et de somme égale à 1, qui pondèrent la contribution de chaque patch au vecteur final. On retrouve ainsi dans cette formule les trois étapes clés : le calcul des scores s_i , leur normalisation en poids a_i par le softmax, et la somme pondérée qui produit z .

3.5.6 Résultats et interprétation

Jeu Hodgkin

Le MIL mono-échelle a été évalué sur le jeu Hodgkin (591 lames : 355 en entraînement, 118 en validation, 118 en test), avec l'encodeur H-optimus-1 (Bioptimus, 2025), successeur plus performant de H-optimus-0. Il atteint une accuracy de 0,932 et une AUC de 0,977. Le détail par classe figure dans le tableau 3.3, les courbes d'apprentissage et la matrice de confusion dans la figure 3.17 en annexe.

TABLE 3.3 – MIL mono-échelle (20×, H-optimus-1) sur le jeu Hodgkin : métriques au niveau lame (test, $n = 118$). Accuracy = 0,932, AUC = 0,977.

Classe	Précision	Rappel	F1	Support
FH	0,939	0,902	0,920	51
Hodgkin	0,928	0,955	0,941	67
Moyenne macro	0,933	0,929	0,931	118
Moyenne pondérée	0,932	0,932	0,932	118

Sur les 118 lames de test, le modèle en classe correctement 110. Parmi les huit erreurs, on compte cinq faux positifs et trois faux négatifs.

Comparaison avec le vote majoritaire. Le tableau 3.4 compare le MIL mono-échelle aux trois approches par vote majoritaire (ResNet50 à 20×, H-optimus-0 à 20× et à 5×) sur le même ensemble de test. À grossissement identique (20×), le passage du vote majoritaire (H-optimus-0) au MIL (H-optimus-1) augmente l'accuracy de 0,864 à 0,932 et l'AUC de 0,956 à 0,977. Ce gain combine l'effet du MIL et celui du changement de version de l'encodeur. Le gain le plus important est porté par le rappel de la classe Hodgkin, qui bondit de 0,821 à 0,955 : le vote majoritaire ratait douze lymphomes sur soixante-sept, contre trois pour le MIL. Ce résultat permet de mettre en évidence l'importance de l'attention : les lymphomes dont la tumeur est concentrée dans un petit nombre de patches se retrouvent en minorité par rapport à la masse de patches qui ne sont pas porteurs d'information.

TABLE 3.4 – Jeu Hodgkin (test, $n = 118$) : comparaison du vote à majorité avec le MIL. Le vote majoritaire décide au niveau patch (chaque patch à poids égal) ; le MIL agrège les patches par attention.

Stratégie (encodeur, échelle)	Accuracy	AUC	Rappel Hodgkin	F1 macro
Vote majoritaire (ResNet50, $20\times$)	0,746	0,866	0,746	0,743
Vote majoritaire (H-optimus-0, $5\times$)	0,856	0,931	0,851	0,854
Vote majoritaire (H-optimus-0, $20\times$)	0,864	0,956	0,821	0,864
MIL mono-échelle (H-optimus-1, $20\times$)	0,932	0,977	0,955	0,931

Instabilité à l'epoch 7. Sur la figure 3.4 ci-dessous, on peut observer un pic à la 7^e epoch. La perte de validation augmente de 0,072 à 0,479 tandis que l'accuracy de validation baisse de 0,983 à 0,797, avant de retrouver dès l'epoch suivante son niveau antérieur (perte 0,043, accuracy 0,975). Ce décrochage ne dure pas et ne change rien au pouvoir discriminant du modèle : la perte d'entraînement reste basse (0,026) et, surtout, l'AUC de validation demeure élevée (0,993).

L'AUC mesure la qualité du classement des lames indépendamment du seuil de décision, alors que l'accuracy dépend du seuil fixé à 0,5. Le maintien d'une AUC quasi parfaite signifie que le modèle classe toujours correctement les lames : la chute d'accuracy vient seulement du fait que beaucoup de lames avaient un score proche de 0,5 et, durant la calibration, ces lames se sont retrouvées de l'autre côté du seuil.

Ce pic est causé par une mise à jour trop brutale de l'optimiseur (AdamW) à ce stade : la perte d'entraînement étant déjà très faible, le modèle évolue dans une région où le paysage de perte est très pentu, où un seul pas peut produire des sorties temporairement surconfiées ; la petite taille de l'ensemble de validation (118 lames) amplifie l'effet observé. L'optimiseur corrige ce dépassement dès l'epoch suivante et, le modèle retenu étant celui de la meilleure epoch de validation, ce pic reste sans conséquence sur les performances rapportées.

La matrice de confusion du MIL mono-échelle (H-optimus-1, $20\times$) est reportée en annexe (figure 3.17).

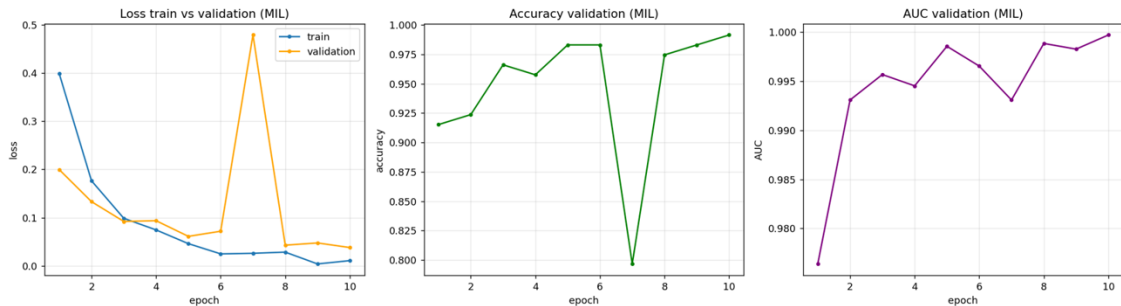


FIGURE 3.4 – Courbes d'apprentissage du MIL mono-échelle ($20\times$, H-optimus-1), jeu Hodgkin.

3.5.7 Exploration multi-échelle en concaténation ($20\times + 5\times$)

À travers mes recherches, j'ai découvert une manière de combiner deux échelles qui cherche à éviter qu'un niveau de résolution ne pénalise l'autre, contrairement à la simple moyenne des scores. Cette stratégie, dite de concaténation, s'inspire des travaux de Li et al. (2021) sur le MIL multi-échelle. Le principe est le suivant. Pour chaque patch, l'encodeur produit un vecteur de features (1536 pour H-optimus-1), qu'une première couche linéaire réduit à 256. Le mécanisme d'attention attribue ensuite à chaque patch un poids : on multiplie le vecteur de chaque patch par son poids, puis on somme les vecteurs de tous les patches. On obtient ainsi un unique vecteur qui représente la lame en insistant sur ses patches les plus informatifs. Le même procédé est appliqué à l'autre résolution, ce qui fournit un second vecteur de lame. Les deux vecteurs sont alors concaténés en un seul vecteur de 512 features, qui rassemble l'information des deux échelles. Ce vecteur traverse enfin un dernier réseau de neurones produisant deux scores, un par classe (bénin et malin). Une fonction softmax les convertit en probabilités comprises entre 0 et 1 et dont la somme vaut 1.

Algorithm 4 : MIL multi-échelle à fusion par concaténation

Input : Les patches d'une lame aux deux échelles : $P_{20\times}$ et $P_{5\times}$

Output : Les deux scores de classe de la lame

```

 $v_{20\times} \leftarrow \text{BrancheAttention}_{20\times}(P_{20\times});$            // vecteur agrégé de l'échelle  $20\times$ 
 $v_{5\times} \leftarrow \text{BrancheAttention}_{5\times}(P_{5\times});$        // vecteur agrégé de l'échelle  $5\times$ 
 $v_{\text{fusion}} \leftarrow [v_{20\times}; v_{5\times}];$                  // concaténation des deux vecteurs
sortie  $\leftarrow \text{Classifieur}(v_{\text{fusion}});$ 
return sortie

```

TABLE 3.5 – MIL multi-échelle par concaténation ($20\times + 5\times$, H-optimus-1) sur le jeu Hodgkin : métriques au niveau lame (test, $n = 118$). Accuracy = 0,949, AUC = 0,984.

Classe	Précision	Rappel	F1	Support
FH	0,959	0,922	0,940	51
Hodgkin	0,942	0,970	0,956	67
Moyenne macro	0,951	0,946	0,948	118
Moyenne pondérée	0,949	0,949	0,949	118

Comparaison avec le MIL mono-échelle. Le tableau 3.6 compare les deux approches, avec le même encodeur (H-optimus-1) et le même ensemble de test. L'ajout du $5\times$ fait légèrement progresser l'accuracy de 0,932 à 0,949 et l'AUC de 0,977 à 0,984 : c'est une différence de deux lames de test. Un écart aussi petit peut être dû à la variabilité et ne représente pas une nette amélioration de la classification.

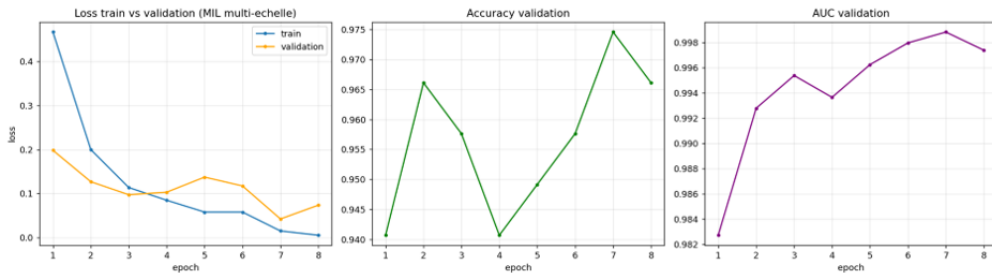
Le bénéfice réellement observable de la concaténation se situe ailleurs : dans la stabilité de l'entraînement. Alors que le modèle mono-échelle présentait un pic marqué à l'époque 7 (effondrement de l'accuracy de validation et explosion de la perte), les courbes du modèle multi-échelle (figure 3.5) restent régulières tout au long de l'apprentissage, sans décrochage de la perte ni de l'accuracy de validation.

TABLE 3.6 – Jeu Hodgkin (test, $n = 118$, encodeur H-optimus-1) : MIL mono-échelle ($20\times$) contre MIL multi-échelle par concaténation ($20\times + 5\times$).

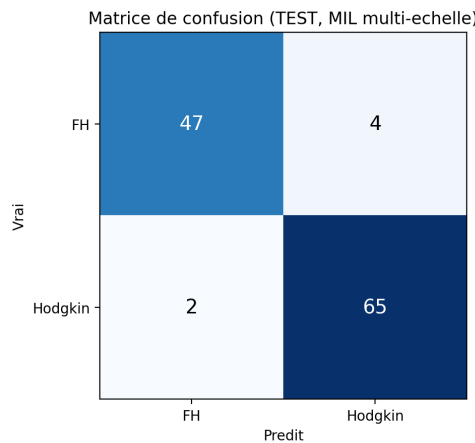
Métrique	MIL mono-échelle ($20\times$)	MIL multi-échelle ($20\times + 5\times$)
Accuracy	0,932	0,949
AUC	0,977	0,984
Rappel FH	0,902	0,922
Rappel Hodgkin	0,955	0,970
F1 macro	0,931	0,948

Cette légère amélioration peut s'expliquer par la conception de H-optimus-1 : c'est un modèle de fondation entraîné à un seul niveau de résolution ($20\times$). Appliqué à des patches en $5\times$, il produit donc des vecteurs de features moins pertinents, car il n'a jamais rencontré cette échelle durant son pré-entraînement. La branche $5\times$ agit alors davantage comme une régularisation, qui lisse et stabilise l'apprentissage, que comme une source d'information complémentaire. À cela s'ajoute le fait que les performances en $20\times$ sont déjà très élevées, ce qui laisse d'autant moins de marge de progression. Ces deux facteurs expliquent un gain presque nul en performance, mais une amélioration de la stabilité.

Afin d'obtenir une réelle plus-value du multi-échelle, un encodeur entraîné à plusieurs grossissements était nécessaire. C'est pourquoi, pour la suite de mon travail, j'ai utilisé Virchow2, un modèle de fondation entraîné à plusieurs grossissements, dont le $20\times$ et le $5\times$.



(a) Courbes d'apprentissage (perte, accuracy et AUC de validation).



(b) Matrice de confusion (test).

FIGURE 3.5 – MIL multi-échelle par concaténation ($20\times + 5\times$, H-optimus-1) sur le jeu Hodgkin.

3.5.8 Améliorations de l'entraînement : early stopping et réduction adaptative du taux d'apprentissage

Une fois l'architecture du modèle fixée, j'ai amélioré la procédure d'entraînement elle-même en y ajoutant deux mécanismes complémentaires : l'arrêt anticipé (early stopping) et la réduction adaptative du taux d'apprentissage (adaptive learning rate).

Le premier répond à un problème classique en apprentissage profond : le surapprentissage. Au fil des epochs, la perte d'entraînement continue de baisser, mais au-delà d'un certain point, le modèle cesse de généraliser et se met à mémoriser les données d'entraînement, si bien que ses performances sur la validation stagnent puis se dégradent. L'early stopping est une solution à ce problème. Il permet de comparer les performances sur le dataset de validation pour chaque nouvelle epoch et conserve le modèle avec les meilleurs résultats. Il stoppe l'entraînement si celui-ci stagne ou baisse puis retient le modèle avec la meilleure validation. Une deuxième solution consiste à ajuster le taux d'apprentissage au cours de l'entraînement. Celui-ci permet de faire baisser la taille des pas de la descente de gradient : si ces pas sont trop grands, le modèle oscille autour de la solution sans vraiment converger et s'ils sont trop petits, l'apprentissage est ralenti.

Un bon compromis consiste à commencer avec un pas relativement grand pour progresser vite, puis à le réduire pour affiner. J'ai donc utilisé un mécanisme qui surveille la perte de validation et divise automatiquement le taux d'apprentissage par deux dès qu'elle cesse de diminuer pendant plusieurs epochs. Le modèle peut ainsi effectuer de petits pas plus précis en fin d'entraînement, une fois arrivé à proximité d'une bonne solution.

Ensemble, ces deux mécanismes rendent l'entraînement à la fois plus robuste et plus économe : la réduction du pas aide le modèle à converger plus finement, tandis que l'arrêt anticipé garantit que l'on conserve sa meilleure version sans gaspiller de calcul.

Algorithm 5 : Entraînement avec réduction adaptative du taux d'apprentissage et arrêt anticipé

Input : Le modèle, les ensembles d'entraînement et de validation, le nombre maximal d'epochs $N_{\max}$, la patience P

Output : Le modèle dans son meilleur état de validation

Initialiser le planificateur du taux d'apprentissage (facteur 0,5, patience 2, minimum 10^{-6});

$meilleur \leftarrow (-\infty, -\infty)$; $sans \leftarrow 0$;

for $ep \leftarrow 1$ **to** $N_{\max}$ **do**

 Entraîner le modèle pendant une epoch;

 Évaluer sur la validation : accuracy a , AUC u et perte ℓ ;

 // Réduction adaptative du pas

if ℓ n'a pas diminué depuis 2 epochs **then**

$\eta \leftarrow \max(\eta/2, 10^{-6})$;

 // Arrêt anticipé : conserver la meilleure version

if $(a, u) > meilleur$ (accuracy d'abord, puis AUC en cas d'égalité) **then**

$meilleur \leftarrow (a, u)$;

 Sauvegarder l'état du modèle et l'epoch ep ;

$sans \leftarrow 0$;

else

$sans \leftarrow sans + 1$;

if $sans \geq P$ **then**

arrêter l'entraînement;

Recharger l'état sauvegardé du modèle;

return le modèle

3.5.9 Cartes d'attention : visualiser où le modèle regarde

Un atout majeur du MIL à attention, au-delà de sa performance, est son interprétabilité : il attribue à chaque patch un poids d'attention a_i . Ces poids, déjà utilisés pour l'agrégation $z = \sum_i a_i h_i$, peuvent être réutilisés pour construire une carte d'attention : on récupère le poids de chaque patch, on le replace à sa position (x, y) sur la lame, et on superpose la carte de chaleur ainsi obtenue à la vignette de la lame. On visualise alors directement les régions sur lesquelles le modèle porte son attention pour prendre sa décision. Le pseudo-code 6 montre les trois étapes que j'ai suivies pour reconstruire cette carte.

Algorithm 6 : Reconstruction et superposition de la carte d'attention

Input : Les features des patches d'une lame, leurs coordonnées, la vignette de la lame
Output : La vignette surmontée de la carte d'attention
 // 1. Poids d'attention de chaque patch
 Passer les features dans le modèle pour obtenir un poids d'attention par patch ;
 // 2. Report de chaque poids à la position du patch
 Initialiser une carte vide aux dimensions de la vignette ;
foreach *patch de coordonnées (x, y) et de poids p* **do**
 Ramener (x, y) à l'échelle de la vignette ;
 Affecter la valeur p à la zone correspondante de la carte ;
 // 3. Superposition
 Afficher la vignette de la lame ;
 Superposer la carte en carte de chaleur semi-transparente ;
 (*bleu = faible attention, rouge = forte attention*) ;
return la vignette annotée

La figure 3.6 en donne un exemple : les patches auxquels le modèle accorde le plus d'importance apparaissent en rouge, les moins déterminants en bleu, sur le fond de la lame originale. On peut ainsi confronter le foyer d'attention du modèle à l'observation d'un pathologiste et vérifier que la décision repose bien sur des zones pertinentes.

Cet outil sera indispensable pour la suite du travail, pour trois raisons. Il permettra d'abord de valider le comportement du modèle, en s'assurant qu'il porte son attention sur des régions cliniquement significatives (follicules, zones d'architecture anormale), le follicule étant la structure sphérique du ganglion où se regroupent les lymphocytes, et dont l'organisation est justement au cœur de la distinction entre hyperplasie et lymphome folliculaires. Il servira ensuite à analyser les erreurs : sur une lame mal classée, observer où l'attention s'est portée aide à distinguer une véritable défaillance du modèle d'un cas intrinsèquement ambigu. Il rendra enfin possible la comparaison des différentes configurations d'encodeurs (H-optimus-1, Virchow2) et d'échelles (branche 20× contre branche 5×), en permettant de voir où leur attention se concentre. La carte d'attention constitue donc le principal levier d'interprétation des résultats présentés dans la suite.

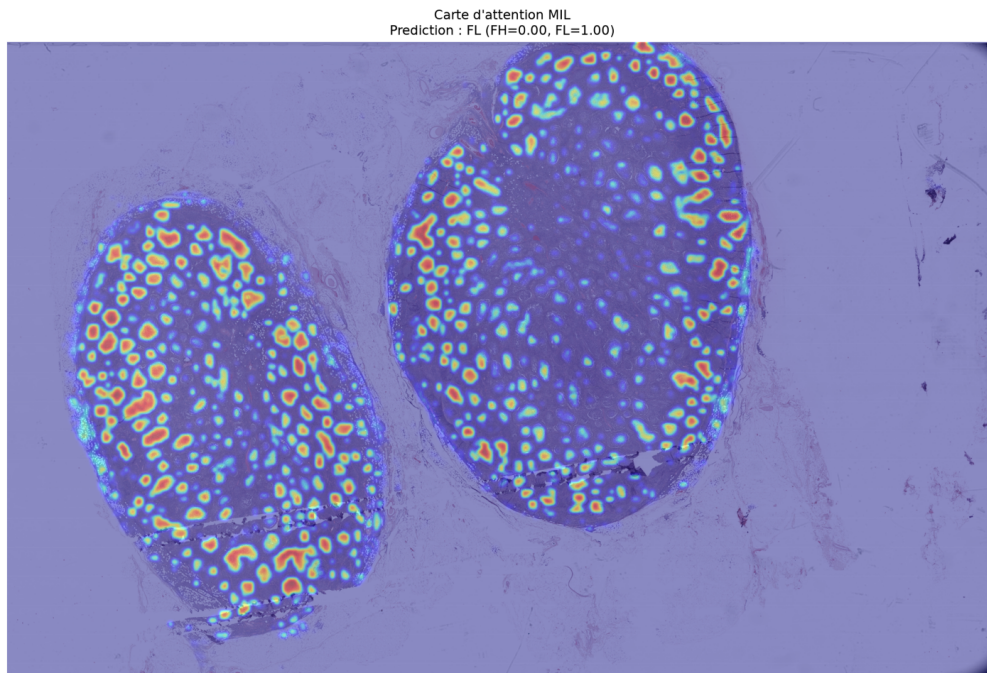


FIGURE 3.6 – Exemple de carte d’attention superposée à une lame : les régions en rouge concentrent l’essentiel du poids de décision du modèle, les régions en bleu n’y contribuent quasiment pas. Dans ce cas, le modèle prédit la classe maligne avec un score de 1, et sa prédiction est juste.

3.6 Phase 2 — Application au lymphome folliculaire

La première phase a permis de valider l’ensemble de la chaîne de traitement sur le jeu Hodgkin. La seconde phase applique cette méthode au problème réellement visé : distinguer une hyperplasie folliculaire (FH, bénigne) d’un lymphome folliculaire (FL, malin). Le point essentiel est la réutilisabilité de la démarche : le pipeline d’extraction des patches et des features, le protocole de découpage (stratifié, sans fuite de patient entre entraînement, validation et test) et l’architecture MIL restent identiques ; seuls les données et les encodeurs évalués changent. Passer d’un problème à l’autre ne demande donc que de fournir de nouvelles lames et d’entraîner un nouveau modèle, ce qui constitue un atout majeur de l’approche.

Il convient de souligner que les performances obtenues sur les deux jeux ne sont pas directement comparables : il s’agit de deux problèmes cliniques distincts, aux difficultés propres, et non de deux évaluations d’une même tâche. Le jeu Hodgkin a joué le rôle de banc de validation méthodologique ; les résultats présentés ci-dessous sont ceux qui portent sur la question cible. Le jeu comprend 651 lames, réparties selon le même protocole en 391 lames d’entraînement, 130 de validation et 130 de test. Deux configurations y sont évaluées : le MIL mono-échelle avec l’encodeur H-optimus-1, et le MIL multi-échelle par concaténation avec l’encodeur Virchow2.

MIL mono-échelle (H-optimus-1, 20×)

Appliqué au jeu lymphome folliculaire avec l’encodeur H-optimus-1, le MIL mono-échelle atteint une accuracy de 0,946 et une AUC de 0,992 sur l’ensemble de test (tableau 3.7). Il laisse échapper cinq lymphomes sur soixante-quinze (rappel FL de 0,933) et seulement deux hyperplasies sont classées à tort en lymphome, ce qui traduit un modèle à la fois sensible et spécifique sur ce problème.

Les courbes d’apprentissage du MIL mono-échelle (H-optimus-1, 20×) sont reportées en annexe (figure 3.18).

TABLE 3.7 – MIL mono-échelle ($20\times$, H-optimus-1) sur le jeu lymphome folliculaire : métriques au niveau lame (test, $n = 130$). Accuracy = 0,946, AUC = 0,992.

Classe	Précision	Rappel	F1	Support
FH	0,914	0,964	0,938	55
FL	0,972	0,933	0,952	75
Moyenne macro	0,943	0,948	0,945	130
Moyenne pondérée	0,948	0,946	0,946	130

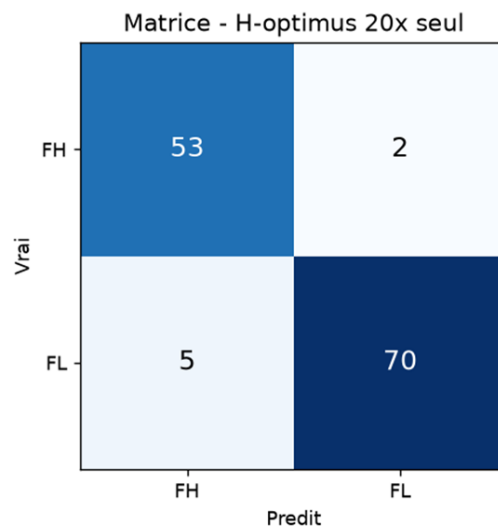


FIGURE 3.7 – Matrice de confusion du MIL mono-échelle ($20\times$, H-optimus-1) sur le jeu lymphome folliculaire.

MIL multi-échelle par concaténation (Virchow2, $20\times + 5\times$)

La configuration multi-échelle reprend le principe de concaténation des features $20\times$ et $5\times$, mais avec l'encodeur Virchow2. Ce choix est motivé par la conclusion tirée sur le jeu Hodgkin : contrairement à H-optimus-1, entraîné au seul grossissement $20\times$, Virchow2 a été entraîné à plusieurs grossissements ($40\times$, $20\times$, $10\times$ et $5\times$). On a donc des raisons de penser qu'il produit de meilleures représentations des patches en $5\times$ qu'un modèle n'ayant jamais rencontré cette échelle, et donc que sa branche $5\times$ apporte une information réellement exploitable.

Un arrêt anticipé (algorithme 5) est utilisé (meilleure epoch retenue : la huitième). Sur l'ensemble de test, ce modèle atteint une accuracy de 0,931 et une AUC de 0,984 (tableau 3.8).

Les courbes d'apprentissage du MIL multi-échelle (Virchow2) sont reportées en annexe (figure 3.19).

TABLE 3.8 – MIL multi-échelle par concaténation ($20\times + 5\times$, Virchow2) sur le jeu lymphome folliculaire : métriques au niveau lame (test, $n = 130$). Accuracy = 0,931, AUC = 0,984.

Classe	Précision	Rappel	F1	Support
FH	0,942	0,891	0,916	55
FL	0,923	0,960	0,941	75
Moyenne macro	0,933	0,925	0,929	130
Moyenne pondérée	0,931	0,931	0,930	130

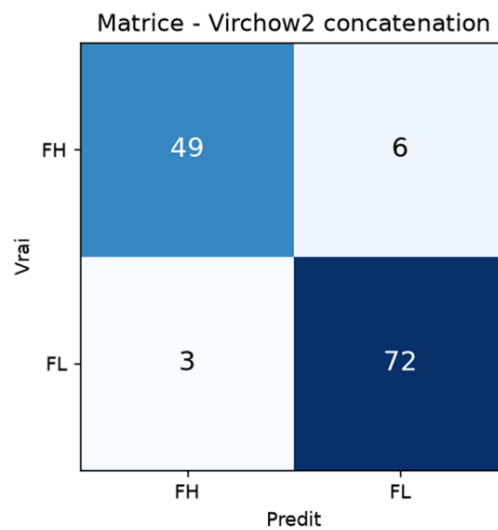


FIGURE 3.8 – Matrice de confusion du MIL multi-échelle ($20\times + 5\times$, Virchow2) sur le jeu lymphome folliculaire.

Comparaison des deux configurations

Les tableaux 3.7 et 3.8 présentent les deux configurations sur le même ensemble de test. Les résultats sont proches et ne désignent pas de vainqueur univoque : le MIL mono-échelle H-optimus-1 obtient la meilleure accuracy (0,946 contre 0,931) ainsi que la meilleure AUC (0,992 contre 0,984), tandis que le MIL multi-échelle Virchow2 obtient un meilleur rappel de la classe FL (0,960 contre 0,933, soit deux lymphomes ratés de moins). Ces écarts doivent toutefois être interprétés avec du recul : ils portent sur un unique entraînement et représentent une différence de deux lames sur les 130 de l'ensemble de test. À cette échelle, le résultat reste sensible au nombre d'époques et à la seed. Le modèle mono-échelle sépare légèrement mieux les deux classes en termes de classement (AUC) et reste le plus performant au seuil de décision de 0,5, mais la version multi-échelle est la plus sûre pour ne pas manquer un lymphome. On remarque par ailleurs que le MIL mono-échelle H-optimus-1 converge en beaucoup moins d'époques (figure 3.18 en annexe) : sa meilleure epoch de validation est atteinte très tôt (entre la deuxième et la cinquième epoch selon les entraînements), là où le MIL multi-échelle Virchow2 continue de progresser bien plus longtemps (meilleure epoch retenue : la huitième, et parfois beaucoup plus tard sur des entraînements prolongés).

Cela peut être expliqué par la qualité des poids internes de H-optimus-1. En effet, ses vecteurs de features permettent déjà de très bien distinguer FH et FL, si bien que le réseau de neurones n’a besoin que de quelques epochs pour se calibrer. Le MIL multi-échelle est à l’inverse plus complexe : il doit apprendre à combiner des vecteurs de features plus grands (2560 dimensions contre 1536), issus de deux échelles de grossissement, ce qui rend l’optimisation beaucoup plus longue. Le modèle mono-échelle est donc aussi plus rapide et plus économique à entraîner. Ce constat, ainsi que l’analyse des cartes d’attention des branches $20\times$ et $5\times$, sont développés dans la partie consacrée à l’analyse des erreurs.

3.6.1 Approche adaptative par double encodeur

Un constat issu de l’analyse des erreurs

L’analyse des erreurs, menée sur plusieurs entraînements en examinant systématiquement les lames mal classées et leurs cartes d’attention, a fait émerger un constat net : le MIL mono-échelle H-optimus-1 ($20\times$) est particulièrement performant sur les biopsies.

Une biopsie est un prélèvement de très petite taille (quelques fragments de tissu prélevés à l’aiguille), par opposition à une pièce chirurgicale, qui correspond à l’exérèse d’un ganglion entier ou d’un fragment étendu — l’exérèse désignant l’ablation chirurgicale de tout ou partie d’un organe ou d’une lésion. Le diagnostic sur biopsie présente une particularité : la tumeur y est de petite taille et concentrée, et sa reconnaissance ne repose pas sur l’architecture tissulaire globale. Le détail cellulaire observable à fort grossissement ($20\times$) suffit alors au diagnostic, tandis que le contexte architectural n’est pas nécessaire, et peut même introduire du bruit.

En revanche, sur les pièces chirurgicales, plus étendues, l’architecture des follicules est bien plus lisible que sur une biopsie : le MIL multi-échelle par concaténation devient alors plus performant, car le grossissement $5\times$ capte cette information architecturale utile au diagnostic.

Illustration sur une biopsie

La lame de la figure 3.9 est un bon exemple de ce phénomène. Il s’agit d’une biopsie de lymphome folliculaire de bas grade : un cas difficile, car le bas grade ressemble à une hyperplasie bénigne et le prélèvement ne contient pas beaucoup de tissu. Les figures 3.9 et 3.10 comparent les cartes d’attention des deux modèles sur cette même lame.

La prédiction du MIL multi-échelle (figure 3.9) est fautive : il prédit FH avec une probabilité de 0,86. C’est un faux négatif, l’erreur la plus grave dans la détection d’une pathologie puisqu’un lymphome n’est pas détecté. Ses cartes d’attention en révèlent la cause. La branche $20\times$ se concentre sur une zone précise et semble détecter une tumeur, tandis que la branche $5\times$ concentre son attention sur quelques points isolés et dispersés. La fusion des deux échelles est ainsi tirée vers une décision erronée par une branche $5\times$ qui ne sait pas où regarder.

Le MIL mono-échelle H-optimus-1 (figure 3.10) prédit au contraire correctement FL, avec une probabilité de 0,93. Sa carte d’attention est révélatrice : l’attention se concentre sur quelques foyers précis et localisés le long des fragments de tissu, exactement les petites zones tumorales concentrées caractéristiques de cette biopsie. En se focalisant sur ces foyers plutôt que sur un contexte architectural absent, le modèle mono-échelle rattrape un lymphome que l’approche multi-échelle avait manqué.

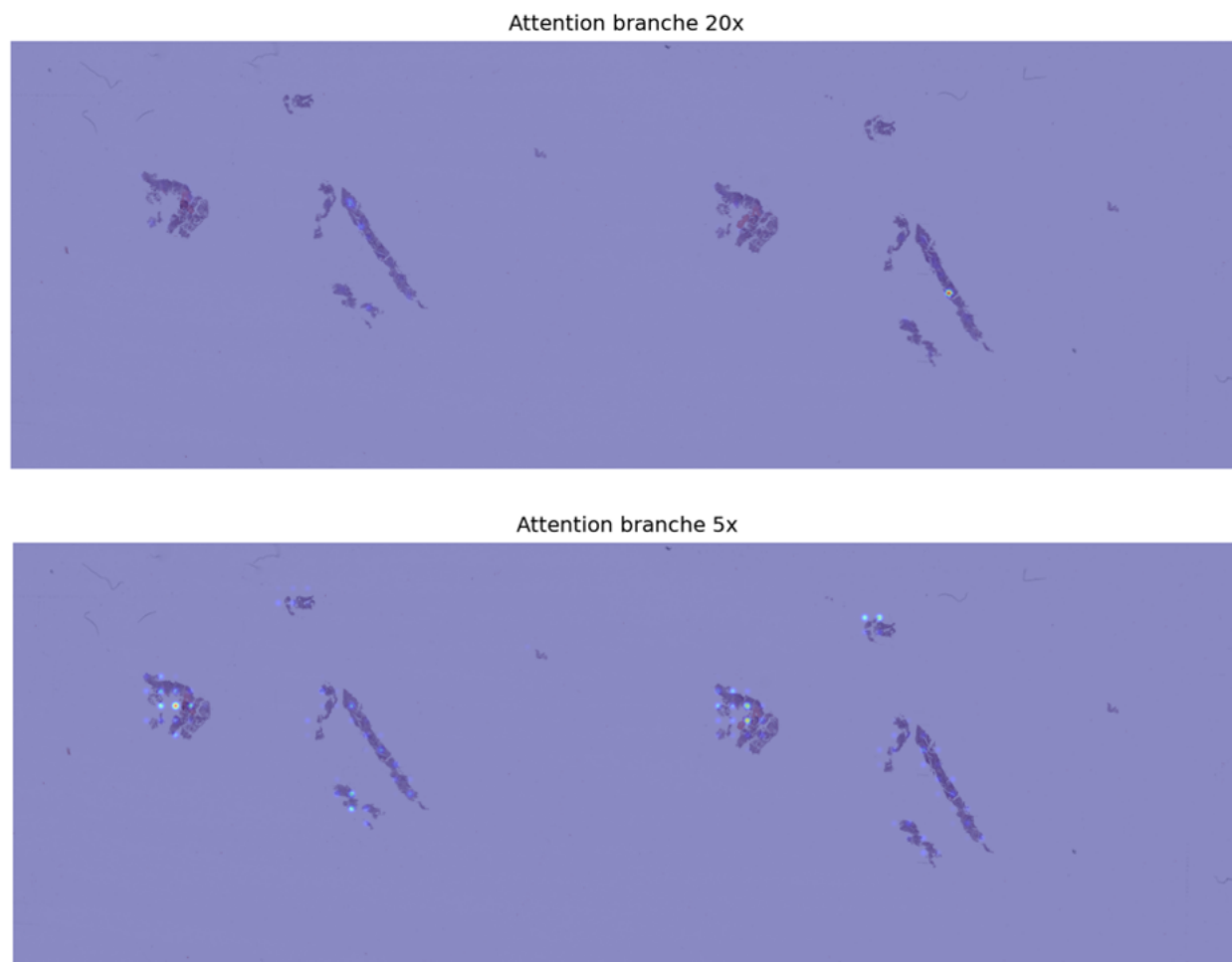


FIGURE 3.9 – Biopsie de lymphome folliculaire de bas grade, MIL multi-échelle (Virchow2, 20× + 5×). Prédiction fausse : FH (0,86). La branche 5× (en bas) porte son attention sur des points isolés et incohérents.

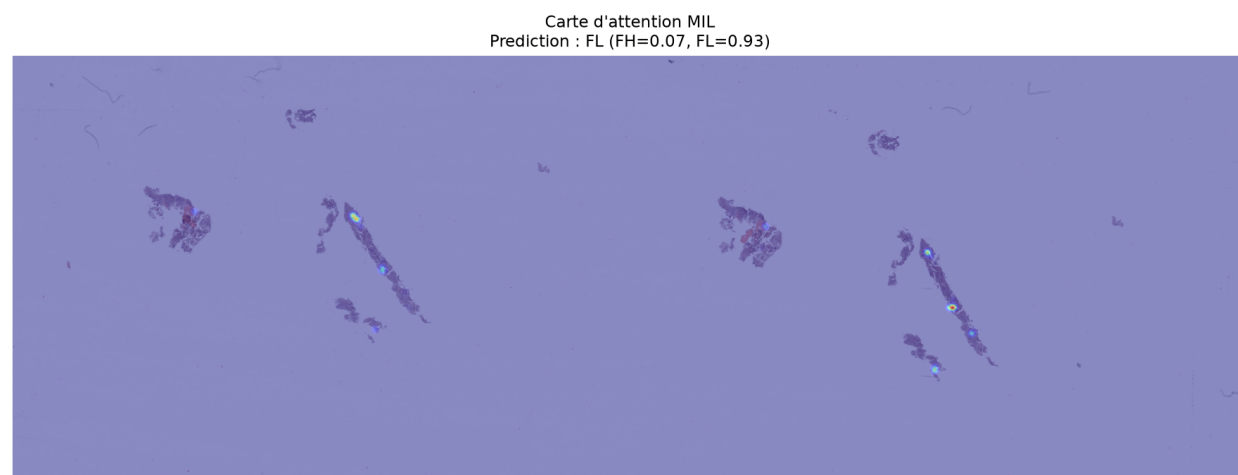


FIGURE 3.10 – Même lame, MIL mono-échelle (H-optimus-1, 20×). Prédiction correcte : FL (0,93). L'attention se concentre sur quelques foyers tumoraux précis.

Le modèle double encodeur

Ce constat m’a permis de développer une approche adaptative : choisir l’encodeur en fonction du type de lame (biopsie ou pièce chirurgicale). Les biopsies sont classées par le MIL mono-échelle H-optimus-1 (20×), bien plus efficace sur ce type de prélèvement, car l’architecture y est peu informative et le 5× n’apporte alors rien de pertinent. Les pièces chirurgicales, plus imposantes, sont traitées par le MIL multi-échelle (Virchow2, 20× + 5×), où le 5× apporte une vraie plus-value diagnostique en captant l’architecture des follicules.

Algorithm 7 : Règle de routage du modèle double encodeur

Input : Les lames du jeu de test avec leur type de prélèvement, le MIL mono-échelle H-optimus-1 (20×), le MIL multi-échelle Virchow2 (20× + 5×)
Output : Une prédiction (FH ou FL) par lame
foreach lame ℓ **du jeu de test** **do**
 if ℓ **est une biopsie** **then**
 // Biopsie : H-optimus-1, 20× seul
 Charger les features H-optimus-1 des patchs 20× de ℓ ;
 $p \leftarrow \text{softmax}(\text{MIL}_{\text{H-optimus-1}}(\text{features } 20\times))$;
 else
 // Pièce chirurgicale : Virchow2, concaténation 20× + 5×
 Charger les features Virchow2 des patchs 20× et 5× de ℓ ;
 $p \leftarrow \text{softmax}(\text{MIL}_{\text{Virchow2}}(\text{features } 20\times, \text{features } 5\times))$;
 Prédire la classe de plus forte probabilité : $\hat{y}_\ell \leftarrow \arg \max_k p_k$;
return les prédictions ($\hat{y}_\ell$)

Résultats

Les trois approches ont été comparées lors d’un même entraînement (100 epochs, patience 30, seed 82), sur le jeu lymphome folliculaire (539 pièces chirurgicales et 112 biopsies), afin d’être directement comparables. Le modèle double encodeur obtient les meilleurs résultats sur la plupart des métriques (tableau 3.9) : une accuracy de 0,962, un rappel FL de 0,973 et un F1 macro de 0,961. Il ne commet que cinq erreurs sur 130 et ne rate que deux lymphomes sur soixante-quinze.

TABLE 3.9 – Jeu lymphome folliculaire (test, $n = 130$, même entraînement, seed 82) : comparaison des trois approches.

Métrique	H-optimus-1 (20×)	Virchow2 (20× + 5×)	Double encodeur
Accuracy	0,946	0,931	0,962
AUC	0,992	0,984	0,989
Rappel FH	0,964	0,891	0,945
Rappel FL	0,933	0,960	0,973
F1 macro	0,945	0,929	0,961
Erreurs (/130)	7	9	5
Lymphomes manqués (/75)	5	3	2

TABLE 3.10 – Double encodeur sur le jeu lymphome folliculaire : métriques au niveau lame (test, $n = 130$). Accuracy = 0,962, AUC = 0,989. Matrice de confusion : vrai FH $\rightarrow$ [52, 3], vrai FL $\rightarrow$ [2, 73].

Classe	Précision	Rappel	F1	Support
FH	0,963	0,945	0,954	55
FL	0,961	0,973	0,967	75
Moyenne macro	0,962	0,959	0,961	130
Moyenne pondérée	0,962	0,962	0,961	130

Analyse des erreurs

La robustesse de l'approche se lit aussi dans la nature de ses erreurs. Grâce au diagnostic précis associé à chaque lame, qui m'a été transmis au cours du stage, l'examen des cinq cas encore mal classés par le double encodeur révèle qu'il ne s'agit jamais d'erreurs simples, mais toujours de situations cliniquement ambiguës.

Trois hyperplasies (FH) sont des faux positifs. La première correspond à une maladie de Castleman, une pathologie qui ne ressemble pas à une hyperplasie classique : le modèle détecte donc une anomalie sur cette lame. Les deux autres, une hyperplasie folliculaire ou nodulaire et un cas d'adénites diverses, sont bénignes mais leur aspect peut ressembler à une prolifération. Sur ces lames, l'ensemble des modèles échoue de la même manière (figure 3.11), ce qui traduit une difficulté intrinsèque et non un défaut propre au double encodeur.

Un échange avec une spécialiste est venu confirmer ce point. Au cours d'une matinée de travail, j'ai pu présenter mes résultats au Dr Charlotte Syrykh, médecin spécialiste en anatomie et cytologie pathologiques, et examiner avec elle les lames sur lesquelles le modèle échouait. Pour la lame de la figure 3.11, mal classée par l'ensemble de mes modèles, le Dr Syrykh m'a confirmé que le diagnostic de référence initial était erroné. Sur cette lame, le modèle ne se trompait donc pas : c'est l'étiquette de référence qui était fausse. Ce constat souligne l'importance d'une relecture experte des cas discordants.

Les deux lymphomes manqués (classés FH, les faux négatifs les plus graves) sont un lymphome folliculaire de grade 1 ou 2 sur ganglion nécrosé (ganglion dont une partie du tissu est morte, ce qui efface son architecture normale) et un lymphome folliculaire de grade 3A. Les lymphomes de bas grade sont précisément ceux qui ressemblent le plus à une hyperplasie bénigne (architecture folliculaire préservée, cellules peu atypiques), et la nécrose complique encore la lecture ; sur la lame nécrosée, le diagnostic prédit varie même d'une epoch à l'autre, signe direct de son ambiguïté. Là encore, ce sont les cas cliniquement les plus limites.

Deux de ces cas (la maladie de Castleman et le lymphome de grade 3A, figure 3.12) étaient d'ailleurs correctement classés par le MIL mono-échelle H-optimus-1, mais, étant des pièces chirurgicales, ils sont aiguillés par le routage vers le MIL multi-échelle qui, lui, se trompe. Cela indique que la règle d'aiguillage actuelle, fondée sur le seul type de prélèvement, n'est pas encore optimale : une règle plus fine, ou une combinaison des deux modèles, pourrait récupérer ces cas et constitue une piste d'amélioration prometteuse.

Au total, les seules erreurs du double encodeur concernent des lames difficiles à interpréter sur le plan pathologique : c'est le signe d'une approche à la fois performante et robuste sur ce problème.

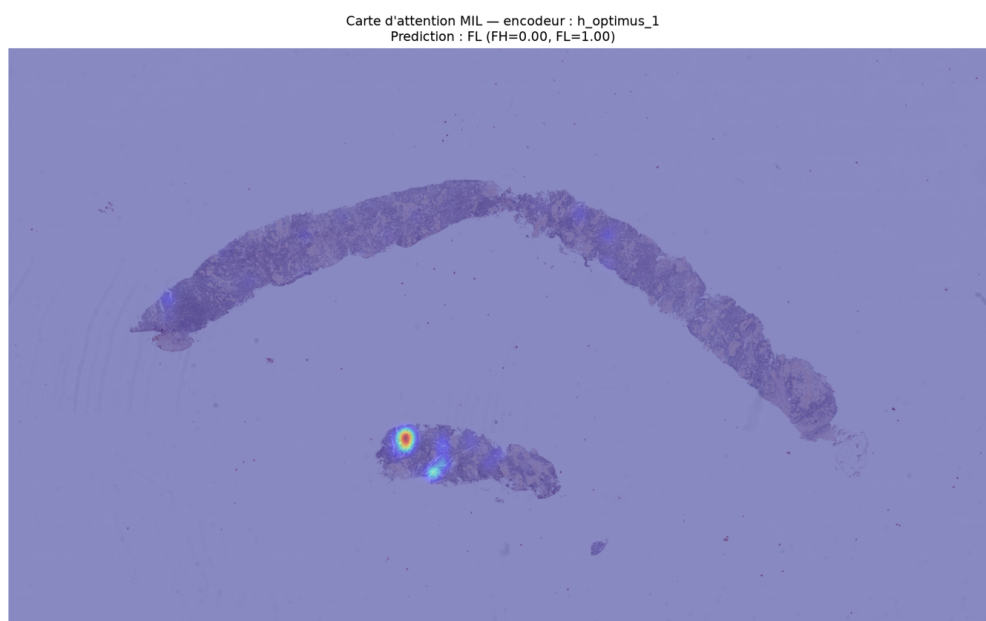


FIGURE 3.11 – Lame étiquetée « adénites diverses » (bénin), dont le diagnostic initial s’est révélé erroné : les trois modèles, qui prédisent FL, avaient en fait raison.

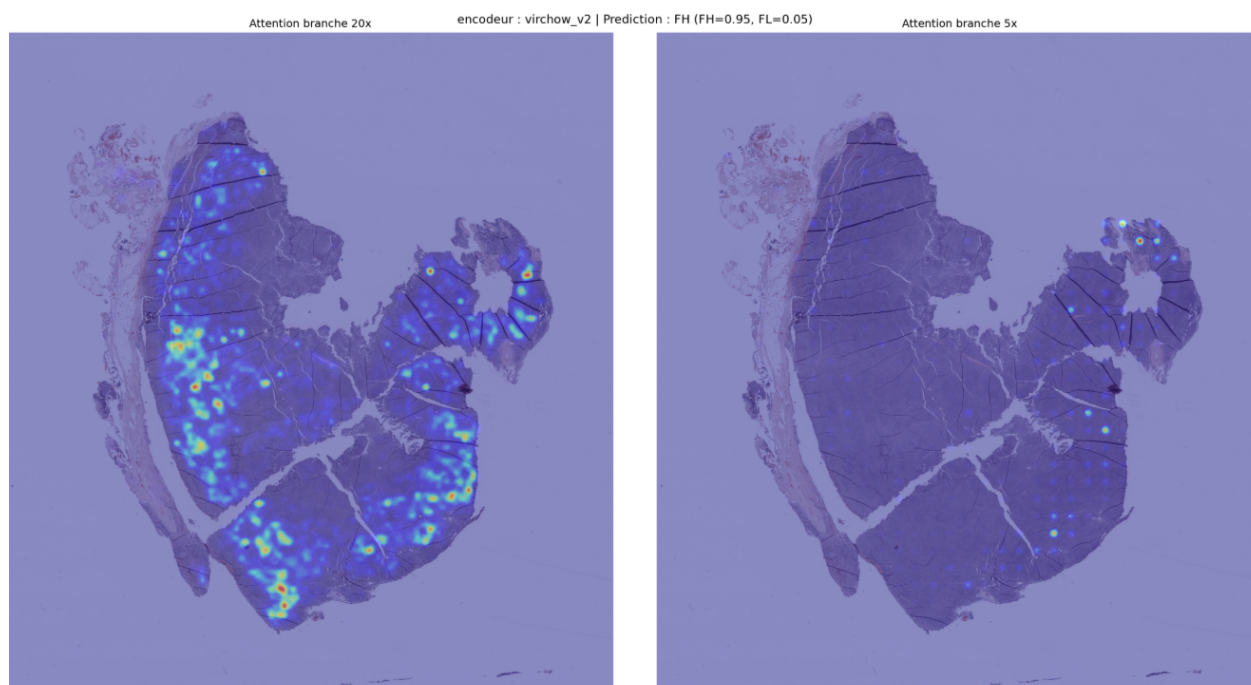


FIGURE 3.12 – Lame dont le diagnostic est un lymphome folliculaire de grade 3A : le MIL multi-échelle prédit à tort FH (0,95), alors que le MIL mono-échelle H-optimus-1 la classait correctement. Cas illustrant la limite de la règle d’aiguillage actuelle.

Limite de l'évaluation. Ce résultat est toutefois à nuancer. L'idée de l'approche adaptative par double encodeur est née de l'analyse des erreurs commises sur un même ensemble de test. Le score de 0,962 est donc à confirmer sur plusieurs ensembles de test, d'autant que les écarts entre modèles ne tiennent qu'à un petit nombre de lames (2/130 entre H-optimus-1 mono-échelle et le double encodeur). Pour affirmer l'intérêt de cette méthode, il faudrait recourir à la validation croisée ou évaluer le modèle sur un jeu de lames indépendant. Par manque de temps, je n'ai pas pu appliquer ces méthodes.

3.7 Pour aller plus loin : quelles dimensions portent l'information ?

Chaque patch est décrit par H-optimus-1 au moyen d'un vecteur de 1536 dimensions. Une question se pose : toutes ces dimensions sont-elles réellement utiles pour distinguer une hyperplasie folliculaire (FH) d'un lymphome folliculaire (FL), ou l'information discriminante est-elle concentrée dans un petit sous-ensemble d'entre elles ? Pour l'explorer, une analyse complémentaire, plus simple que le MIL, a été menée au niveau lame.

Démarche

Chaque lame du jeu lymphome folliculaire est d'abord résumée par un unique vecteur, obtenu en moyennant les features de tous ses patches ; on obtient ainsi une matrice tabulaire de 651×1536 (lames $\times$ dimensions). Les vecteurs sont normalisés (moyenne nulle, variance unité) en n'ajustant les paramètres que sur l'ensemble d'entraînement, afin d'éviter toute fuite du test. Deux méthodes de sélection sont alors appliquées : un test F univarié, qui classe les dimensions selon leur corrélation au label, et une régression logistique pénalisée en L_1 (Lasso), qui met naturellement à zéro les dimensions inutiles. Enfin, on mesure l'AUC obtenue par une régression logistique en validation croisée par patient, en faisant varier le nombre de dimensions conservées (algorithme 8). La normalisation et la sélection y sont refaites dans chaque pli, sur les seuls plis d'entraînement, afin que le pli d'évaluation ne serve jamais à choisir les dimensions.

Algorithm 8 : AUC en validation croisée par patient selon le nombre de features conservées

Input : Features d'entraînement X_{tr} , labels y_{tr} , patients g_{tr}

Output : L'AUC obtenue pour chaque nombre de dimensions conservées

Découper X_{tr} en 5 plis stratifiés et groupés par patient ;

foreach nombre de dimensions k à tester **do**

foreach pli f **do**

 Ajuster la normalisation (moyenne nulle, variance unité) sur les 4 autres plis uniquement ;

if k est inférieur au nombre total de dimensions **then** choisir les k meilleures dimensions par le test F , sur ces 4 plis uniquement ;

 Entraîner une régression logistique sur ces 4 plis, puis mesurer l'AUC sur le pli f ;

 AUC de $k \leftarrow$ moyenne des 5 AUC ;

return les AUC obtenues pour chaque valeur de k

TABLE 3.11 – AUC moyenne (validation croisée par patient) selon le nombre de features H-optimus-1 conservées.

Features	10	25	50	100	200	500	1536
AUC	0,945	0,976	0,975	0,982	0,983	0,985	0,982
Écart-type	$\pm 0,007$	$\pm 0,013$	$\pm 0,009$	$\pm 0,008$	$\pm 0,006$	$\pm 0,008$	$\pm 0,009$

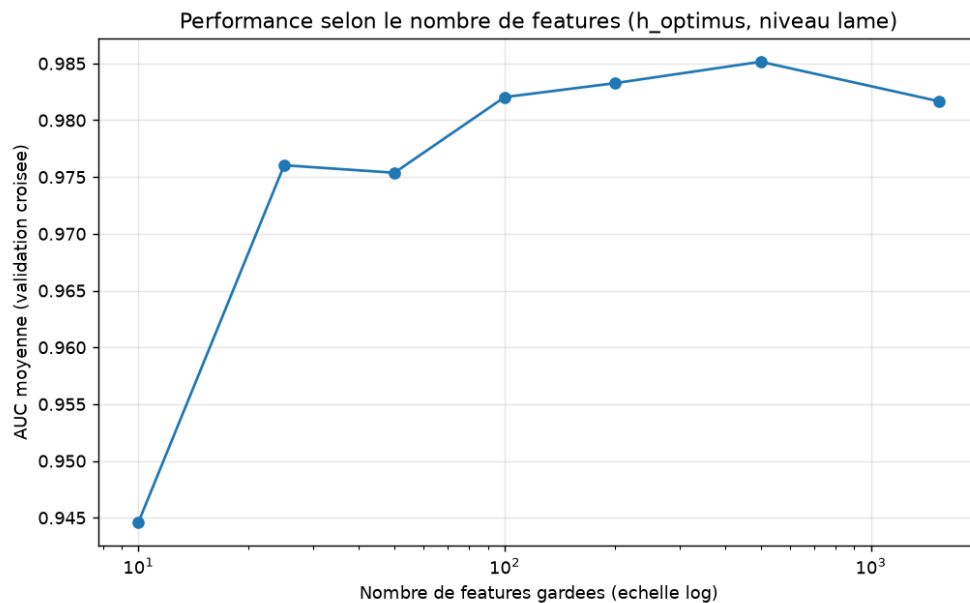


FIGURE 3.13 – AUC moyenne (validation croisée par patient) en fonction du nombre de features H-optimus-1 conservées (échelle logarithmique). La performance est déjà élevée avec 10 dimensions et sature à partir d'une centaine.

Résultats

Avec seulement 10 dimensions sur 1536, soit moins de 1 % d'entre elles, la régression logistique atteint déjà une AUC de 0,945 en validation croisée (tableau 3.11, figure 3.13). Dès 25 dimensions, l'AUC monte à 0,976, et à partir d'une centaine elle rejoint (0,982) celle obtenue avec la totalité des 1536 dimensions ; le meilleur score est même atteint autour de 500 dimensions (0,985), légèrement au-dessus du vecteur complet. On peut en déduire qu'ajouter des dimensions au-delà d'une centaine n'apporte quasiment rien. De son côté, la régression L_1 ne conserve que 40 dimensions non nulles sur 1536 pour obtenir un modèle performant.

Interprétation

Cette analyse de dimension permet de montrer que l'information utile à la distinction FH/FL est concentrée dans un petit nombre de dimensions des vecteurs produits par H-optimus-1 : la grande majorité d'entre elles ne sont pas utiles. Ce résultat ouvre plusieurs perspectives. La première serait l'exploitation et la réduction de ces dimensions, qui permettrait de créer des modèles plus légers et plus rapides. De plus, l'identification et la compréhension des dimensions déterminantes constitueraient un pas vers la compréhension de ce que le modèle regarde.

3.8 Discussion et comparaison avec la littérature

Cette partie met en perspective les méthodes et les résultats obtenus au cours du stage avec les travaux existants, en particulier ceux menés au sein même du laboratoire d'anatomopathologie de l'IUCT-Oncopole.

3.8.1 Comparaison avec les travaux de Syrykh *et al.* (2020)

En 2020, l'équipe de l'IUCT-Oncopole a publié une étude portant sur la même problématique : distinguer automatiquement un lymphome folliculaire d'une hyperplasie folliculaire à partir de lames colorées à l'H&E (Syrykh *et al.*, 2020). Leur jeu de données comptait 378 ganglions (197 lymphomes folliculaires et 181 hyperplasies), issus de deux centres (Toulouse et Dijon), contre 651 lames dans mon travail.

Les méthodes sont différentes sur plusieurs aspects, ce qui reflète l'évolution rapide du domaine en cinq ans :

- **Apprentissage des caractéristiques.** Syrykh *et al.* ont entraîné des réseaux convolutifs directement sur les patches, chaque patch héritant de l'étiquette de sa lame. Dans mon pipeline, les caractéristiques sont extraites par des modèles de fondation déjà existants (H-optimus-1, Virchow2), qui sont des transformers : une architecture plus récente et généralement plus performante, pré-entraînée sur de très grands volumes de lames.
- **Agrégation à l'échelle de la lame.** Dans l'étude de 2020, le diagnostic d'une lame est obtenu en faisant la moyenne des prédictions de tous ses patches, chacun ayant le même poids. Le MIL à attention apprend au contraire à pondérer les patches selon leur pertinence, ce qui permet de se concentrer sur les régions avec de l'information, et de produire des cartes d'attention interprétables. C'est le passage du vote majoritaire au MIL à attention qui a apporté le gain le plus important au cours de mon stage.
- **Estimation de l'incertitude.** L'apport principal de l'étude de 2020 était l'utilisation de réseaux bayésiens, capables d'associer à chaque prédiction un niveau de certitude et de signaler les cas difficiles ou inhabituels. Mon modèle ne dispose pas de ce mécanisme.

Sur le plan des résultats, Syrykh *et al.* obtenaient une AUC comprise entre 0,92 et 0,99 selon la résolution utilisée, tandis que mon modèle à double encodeur atteint une AUC de 0,989 et une accuracy de 0,962 (5 erreurs sur 130 lames de test). Ces chiffres, bien que similaires, ne sont toutefois pas directement comparables. En plus des différences de jeux de données, les protocoles de découpage sont différents : dans l'étude de 2020, les lames ont été réparties aléatoirement entre les ensembles d'entraînement, de validation et de test, alors que j'ai effectué un découpage par patient. Ce dernier garantit que des lames d'un même patient ne se retrouvent jamais à la fois dans l'entraînement et dans le test, ce qui évite que le modèle s'appuie sur des caractéristiques propres à un patient plutôt qu'à la pathologie. Ce type de découpage, devenu depuis une pratique recommandée en pathologie numérique, conduit généralement à une évaluation plus exigeante.

Enfin, les auteurs suggéraient que leur approche pourrait aussi s'appliquer à des prélèvements de petite taille comme les biopsies. Mon travail apporte un élément de réponse à cette question : biopsies et pièces chirurgicales sont différentes d'un point de vue architectural, et les traiter avec des modèles distincts améliore les résultats.

3.8.2 Fusion multi-échelle par concaténation

Le principe de combiner plusieurs grossissements existait déjà. Campanella et al. (2019) combinaient des modèles entraînés à différents grossissements par fusion tardive, et Li et al. (2021) concaténaient les caractéristiques d'un même patch extraites à plusieurs grossissements avant l'agrégation MIL, une approche proche de celle que j'ai utilisée.

Les résultats de la littérature sur l'intérêt du multi-échelle pour les lymphomes sont toutefois nuancés. Une étude comparative récente portant sur le sous-typage des lymphomes (lymphome folliculaire, LLC, lymphome du manteau, LDGCB) n'a observé aucun gain à combiner plusieurs grossissements (Umer et al., 2025). Cette étude portait cependant sur des lames décrites comme des biopsies ganglionnaires. Ce résultat est donc cohérent avec le mien : je n'observe pas non plus de gain sur les biopsies, mais seulement sur les pièces chirurgicales.

3.8.3 Approche à double encodeur

La complémentarité entre modèles de fondation a déjà été prouvée : Neidlinger et al. (2025) ont montré, sur une étude comparative de 19 modèles de fondation, que des modèles entraînés sur des données différentes apprennent des caractéristiques complémentaires, et que leur combinaison surpasse souvent chaque modèle pris isolément. Plus largement, plusieurs études comparatives montrent qu'aucun modèle de fondation n'est le plus performant sur toutes les tâches : chacun excelle sur certains types de problèmes mais pas sur d'autres (Ma et al., 2024). Cette variabilité a conduit à des approches qui adaptent l'utilisation des modèles à chaque tâche plutôt que de s'appuyer sur un modèle unique (Lei et al., 2025). L'approche que j'ai développée utilise cette logique, mais de manière simplifiée : au lieu de fusionner les modèles, elle utilise le modèle le plus adapté en fonction du type de lames (les biopsies vers le modèle 20 $\times$, les pièces chirurgicales vers le modèle multi-échelle). À ma connaissance, dans les travaux que j'ai consultés, aucune étude portant sur la distinction entre hyperplasie et lymphome folliculaire n'utilise une telle approche.

3.9 Difficultés rencontrées

La mise en œuvre de ce pipeline s'est heurtée à plusieurs contraintes matérielles et techniques, principalement liées au volume considérable des données.

La première difficulté a été d'ordre matériel. Les lames numérisées étaient initialement stockées sur une clé USB, qui avait tendance à surchauffer lors des accès prolongés. Cette surchauffe ralentissait les lectures et rendait certaines opérations instables sur de longues durées : je devais interrompre le programme, attendre que la clé refroidisse, puis relancer le traitement. Ces interruptions répétées étaient difficilement compatibles avec des étapes de calcul longues et continues comme l'encodage et le découpage en patches. J'ai donc fini par transférer l'ensemble des données sur le disque de l'ordinateur, ce qui a résolu le problème de surchauffe.

Ce transfert a par contre mis en évidence une contrainte de stockage. Comme évoqué précédemment, les lames numérisées sont extrêmement lourdes : les centaines de lames du projet représentaient à elles seules environ un téraoctet de stockage, auquel sont venus s'ajouter les patches extraits. J'ai donc été confronté à des problèmes de manque d'espace disque, qui m'ont obligé à gérer les données avec précaution : suppression des fichiers intermédiaires et répartition sur plusieurs disques.

La deuxième difficulté majeure a été le temps de calcul. L'encodage des lames par les modèles de fondation est une étape particulièrement chronophage : selon l'encodeur et l'échelle, un encodage complet a pu prendre jusqu'à trois jours (environ 75 heures). Les entraînements eux-mêmes étaient également longs, en particulier pour les derniers modèles, qui comportaient de nombreuses epochs et pouvaient nécessiter entre six et huit heures d'exécution.

Ces contraintes expliquent en partie certains choix faits au cours du stage, comme le plafonnement du nombre de patches par lame ou l'ajout de l'arrêt anticipé, qui ont permis de limiter les temps de calcul sans dégrader significativement les performances.

Conclusion et perspectives

Ce stage m'a permis de construire, du début à la fin, un pipeline complet de pathologie computationnelle destiné à distinguer une hyperplasie folliculaire d'un lymphome folliculaire sur lames numériques. J'ai d'abord mis au point ma méthode sur un premier jeu de données (lymphome de Hodgkin), en commençant par le découpage des lames en patches, leur encodage par des modèles de fondation, un découpage des données par patient, puis une classification par apprentissage à instances multiples à attention. Mes modèles sont devenus de plus en plus performants au fil du stage, du vote majoritaire avec ResNet50 jusqu'au MIL avec H-optimus-1 et Virchow2. J'ai ensuite réutilisé mon pipeline sur le lymphome folliculaire et conçu un modèle à double encodeur qui s'adapte au type de lame pour être le plus efficace possible. Celui-ci a atteint d'excellents résultats, avec une accuracy de 0,962 et une AUC de 0,989, ne se trompant que sur des cas cliniquement ambigus et détectant même une lame dont le diagnostic initial était erroné. Au-delà de ces résultats, j'ai développé un modèle réutilisable en fonction du problème, sans modifier la chaîne de traitement, ce qui est un résultat d'ingénierie à part entière.

Plusieurs perspectives se dégagent de ce travail. En l'état actuel, le pipeline constitue un outil d'aide au diagnostic destiné aux médecins : il ne remplace pas le pathologiste mais lui fournit un second avis objectif, et utile sur les cas ambigus, où réside justement la difficulté. Pour qu'un tel outil soit réellement exploitable par tous, et non par les seuls utilisateurs à l'aise avec le code, il manque encore une application dotée d'une interface où il suffirait de déposer une lame numérique pour obtenir un diagnostic et son degré de confiance. Ce travail s'inscrit plus largement dans la thèse de Stéphane Treillard, qui porte sur la classification par IA des lymphomes.

Ce stage a également été la découverte d'un monde qui m'était inconnu, celui de la recherche médicale et de l'anatomopathologie. Il m'a montré la transversalité de ce type de projet : c'est en examinant avec le Dr Syrykh les lames mal classées que j'ai compris ce que révélaient les erreurs de mon modèle. Plus généralement, l'expertise apportée par les médecins m'a guidé tout au long de mon stage.

Être au contact des chercheurs et des médecins, discuter des problématiques et apporter des éléments de réponse avec mes modèles m'a passionné. J'ai eu le sentiment, à ma petite échelle de stagiaire, de servir à quelque chose d'utile et de faire partie d'un projet porteur de sens.

Ce stage a éveillé en moi une nouvelle envie : celle d'orienter ma future carrière dans le domaine de l'intelligence artificielle appliquée à la médecine. Elle est dans la continuité de mon cursus d'ingénieur en Systèmes Embarqués et Télécommunications, dont les compétences en programmation, intelligence artificielle et traitement d'images se sont révélées indispensables pour ce type de problème. Je souhaite désormais poursuivre dans cette voie, à l'intersection de l'ingénierie et de la santé, un domaine pour lequel je me suis découvert autant d'intérêt que de motivation.

Annexes

Annexe 1

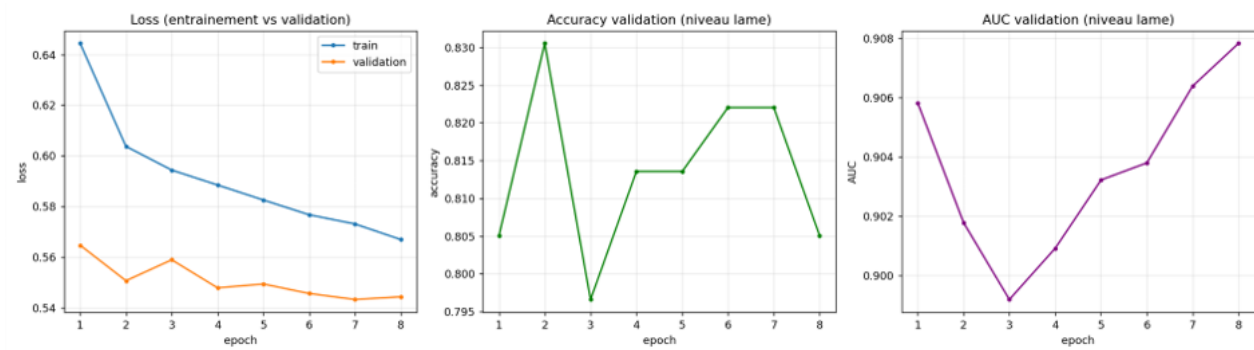


FIGURE 3.14 – Courbes d'apprentissage avec ResNet50 (20×) en vote majoritaire, jeu Hodgkin : perte (entraînement et validation), accuracy et AUC de validation.

Annexe 2

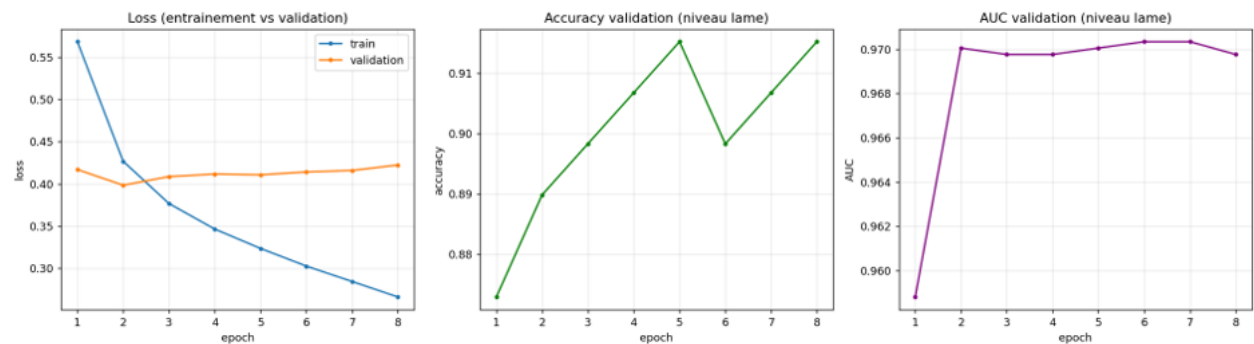
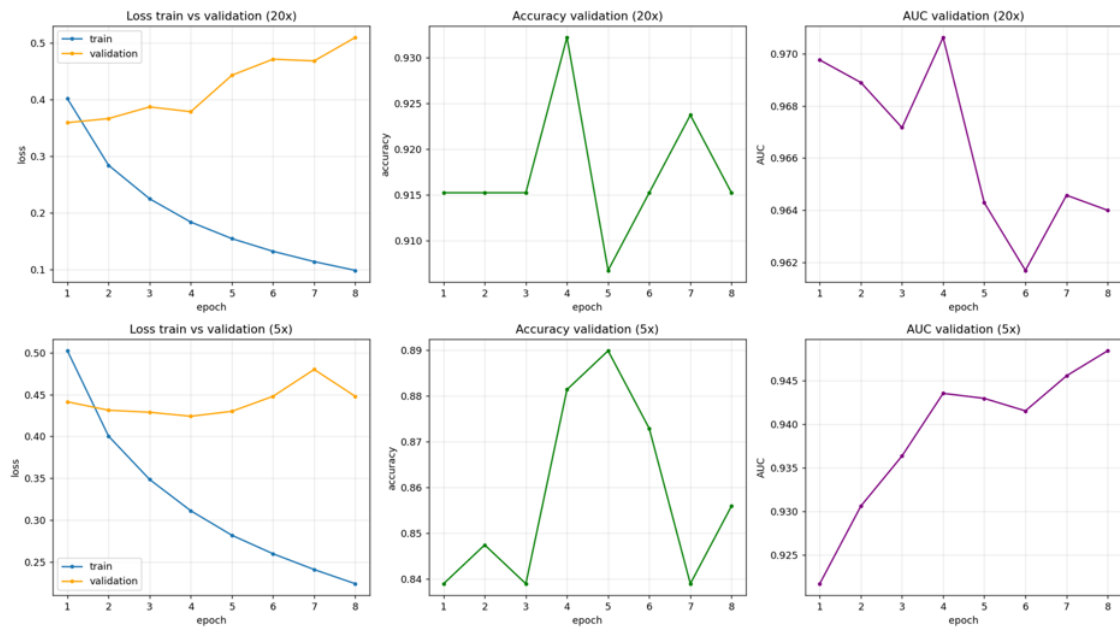
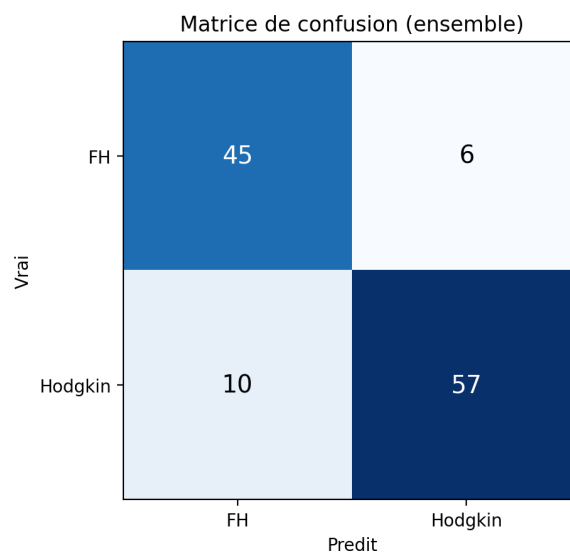


FIGURE 3.15 – Courbes d'apprentissage avec H-optimus-0 (20×) en vote majoritaire, jeu Hodgkin : perte (entraînement et validation), accuracy et AUC de validation.

Annexe 3



(a) Courbes d'apprentissage.



(b) Matrice de confusion (test).

FIGURE 3.16 – Vote majoritaire et moyenne des scores $5\times + 20\times$ avec H-optimus-0, jeu Hodgkin.

Annexe 4

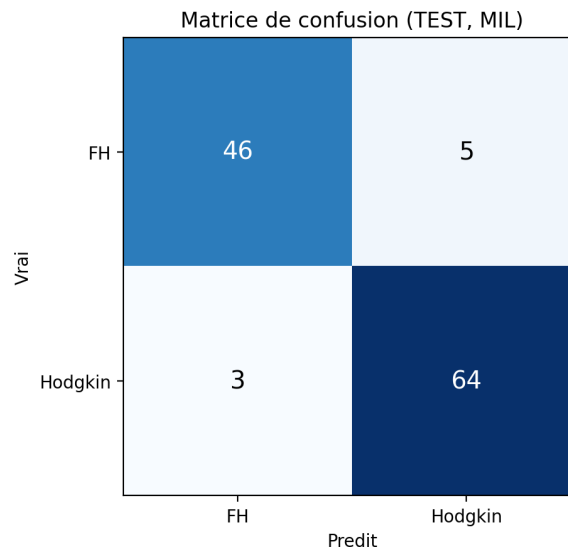


FIGURE 3.17 – Matrice de confusion (test) du MIL mono-échelle ($20\times$, H-optimus-1), jeu Hodgkin.

Annexe 5

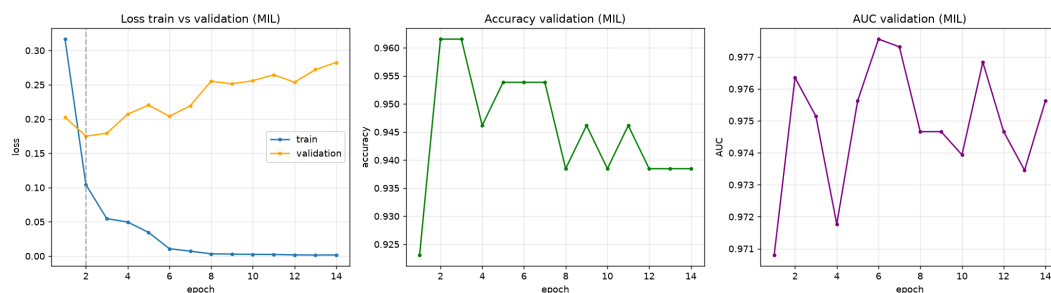


FIGURE 3.18 – Courbes d'apprentissage du MIL mono-échelle ($20\times$, H-optimus-1), jeu lymphome folliculaire.

Annexe 6

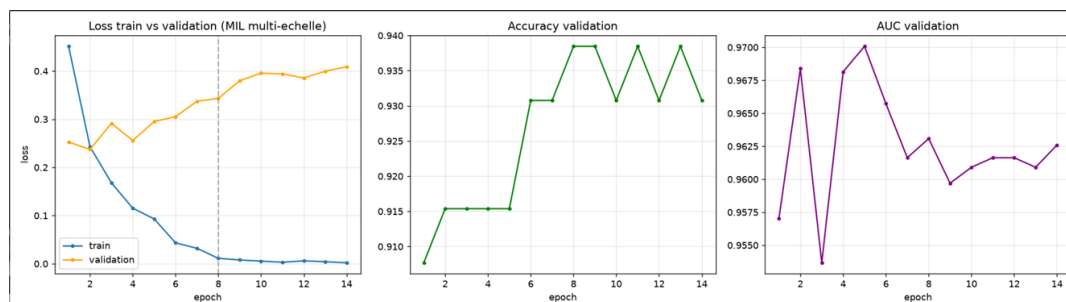


FIGURE 3.19 – Courbes d'apprentissage du MIL multi-échelle ($20\times + 5\times$, Virchow2), jeu lymphome folliculaire.

Bibliographie / Webographie

- L’Agenda 2030 en France, « Vers un modèle hospitalier durable : l’exemple du CHU de Toulouse », 6 mai 2026. En ligne : <https://www.agenda-2030.fr/a-la-une/actualites-a-la-une/article/vers-un-modele-hospitalier-durable-l-exemple-du-chu-de-toulouse>. Consulté le 29 septembre 2026.
- Alagha A., Leclerc C., Kotp Y. et al., « AtlasPatch : Efficient Tissue Detection and High-throughput Patch Extraction for Computational Pathology at Scale », arXiv preprint arXiv :2602.03998, 2026. En ligne : <https://arxiv.org/abs/2602.03998>. Consulté le 6 août 2026.
- Bioptimus, « H-optimus-0 », modèle de fondation en pathologie, 2024. En ligne : <https://huggingface.co/bioptimus/H-optimus-0>. Consulté le 10 septembre 2026.
- Bioptimus, « H-optimus-1 », modèle de fondation en pathologie, 2025. En ligne : <https://huggingface.co/bioptimus/H-optimus-1>. Consulté le 25 août 2026.
- Campanella G., Hanna M. G., Geneslaw L. et al., « Clinical-grade computational pathology using weakly supervised deep learning on whole slide images », *Nature Medicine*, vol. 25, n° 8, p. 1301–1309, 2019. DOI : 10.1038/s41591-019-0508-1. Consulté le 2 septembre 2026.
- CHU de Toulouse, « Politique RSE » et rapport annuel RSE 2024. En ligne : <https://www.chu-toulouse.fr/politique-rse>. Consulté le 29 septembre 2026.
- CHU de Toulouse, « Semaine européenne du développement durable : venez découvrir dans nos établissements les initiatives du CHU ! », 2022. En ligne : <https://www.chu-toulouse.fr/semaine-europeenne-du-developpement-durable-venez>. Consulté le 29 septembre 2026.
- Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., « ImageNet : A Large-Scale Hierarchical Image Database », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 248–255, 2009. DOI : 10.1109/CVPR.2009.5206848. Consulté le 8 septembre 2026.
- Dosovitskiy A., Beyer L., Kolesnikov A. et al., « An Image Is Worth 16×16 Words : Transformers for Image Recognition at Scale », *International Conference on Learning Representations (ICLR)*, 2021. arXiv :2010.11929. Consulté le 12 septembre 2026.
- He K., Zhang X., Ren S., Sun J., « Deep Residual Learning for Image Recognition », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778, 2016. DOI : 10.1109/CVPR.2016.90. Consulté le 1^{er} octobre 2026.
- IBiSA, « Imag’IN », fiche de la plateforme de pathologie numérique de l’IUCT-Oncopole, mise à jour en 2021. En ligne : <https://www.ibisa.net/plateformes/imag-in-368.html>. Consulté le 8 octobre 2026.
- Ilse M., Tomczak J. M., Welling M., « Attention-based Deep Multiple Instance Learning », *Proceedings of the 35th International Conference on Machine Learning (ICML)*, PMLR 80, p. 2127–2136, 2018. Consulté le 13 septembre 2026.
- IUCT-Oncopole, « Anatomopathology », page du département d’anatomie et cytologie pathologiques. En ligne : <https://www.iuct-oncopole.fr/en/anatomical-pathology-department>. Consulté le 8 octobre 2026.
- IUCT-Oncopole, « APRIORICS : l’intelligence artificielle au service d’une médecine personnalisée dans le cancer du sein », communiqué de presse, 10 août 2020. En ligne : <https://www.iuct-oncopole.fr/documents/20049/141329/CP+Apriorics+ao%C3%BBt+2020/c7937eb0-c0e4-4fa0-b105-9606c6254a34>. Consulté le 8 octobre 2026.

- IUCT-Oncopole, « Rapport d'activité 2023 », 2024. En ligne : <https://www.iuct-oncopole.fr/documents/2110567/0/IUCT-Oncopole-RA2023+Digital+FR+bd.pdf/fddc5b5a-53f9-4744-a775-8993599d93d2>. Consulté le 8 octobre 2026.
- Lei W., Li A., Tan Y., Chen H., Zhang X., « Shazam : Unifying Multiple Foundation Models for Advanced Computational Pathology », arXiv preprint arXiv :2503.00736, 2025. Consulté le 6 octobre 2026.
- Li B., Li Y., Eliceiri K. W., « Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 14318–14328, 2021. Consulté le 15 septembre 2026.
- Ma J., Guo Z., Zhou F. et al., « Towards a Generalizable Pathology Foundation Model via Unified Knowledge Distillation », arXiv preprint arXiv :2407.18449, 2024. Consulté le 6 octobre 2026.
- Neidlinger P., El Nahhas O. S. M., Muti H. S. et al., « Benchmarking foundation models as feature extractors for weakly supervised computational pathology », *Nature Biomedical Engineering*, 2025. DOI : 10.1038/s41551-025-01516-3. Consulté le 5 octobre 2026.
- Ravi N., Gabeur V., Hu Y.-T., Hu R., Ryali C., Ma T., Khedr H., Rädle R., Rolland C., Gustafson L., Mintun E., Pan J., Alwala K. V., Carion N., Wu C.-Y., Girshick R., Dollár P., Feichtenhofer C., « SAM 2 : Segment Anything in Images and Videos », *arXiv preprint*, arXiv :2408.00714, 2024. DOI : 10.48550/arXiv.2408.00714. Consulté le 19 septembre 2026.
- Syrykh C., Abreu A., Amara N. et al., « Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning », *npj Digital Medicine*, vol. 3, n° 63, 2020. DOI : 10.1038/s41746-020-0272-0. Consulté le 5 octobre 2026.
- TecHopital, « Développement durable et hôpitaux, le CHU de Toulouse socialement responsable », entretien avec É. Toutut, responsable du développement durable du CHU. En ligne : <https://www.elisabeth-toutut-picard.fr/developpement-durable-hopitaux-chu-de-toulouse-socialement-responsable/>. Consulté le 29 septembre 2026.
- Umer R. M., Sens D., Noll J. et al., « A Multicenter Benchmark of Multiple Instance Learning Models for Lymphoma Subtyping from HE-stained Whole Slide Images », arXiv preprint arXiv :2512.14640, 2025. Consulté le 5 octobre 2026.
- Unicancer, « IUCT Oncopole – Institut Claudius Regaud », présentation du réseau des Centres de lutte contre le cancer. En ligne : <https://www.unicancer.fr/en/clcc/iuct-oncopole-institut-claudius-regaud/>. Consulté le 8 octobre 2026.
- Wang Y., Williamson K. E., Kelly P. J., James J. A., Hamilton P. W., « SurfaceSlide : A Multitouch Digital Pathology Platform », *PLoS ONE*, vol. 7, n° 1, e30783, 2012. DOI : 10.1371/journal.pone.0030783. Consulté le 20 août 2026.
- Zimmermann E., Vorontsov E., Viret J. et al., « Virchow2 : Scaling Self-Supervised Mixed Magnification Models in Pathology », arXiv preprint arXiv :2408.00738, 2024. En ligne : <https://arxiv.org/abs/2408.00738>. Consulté le 8 septembre 2026.

Liste des sigles et abréviations

ACP	Anatomie et Cytologie Pathologiques
AUC	<i>Area Under the (ROC) Curve</i> — aire sous la courbe ROC
CHU	Centre Hospitalier Universitaire
CLCC	Centre de Lutte Contre le Cancer
CNN	<i>Convolutional Neural Network</i> — réseau de neurones convolutif
CRCT	Centre de Recherches en Cancérologie de Toulouse
CSE	Comité Social et Économique
FFPE	<i>Formalin-Fixed, Paraffin-Embedded</i> — tissu fixé et inclus en paraffine
FH	<i>Follicular Hyperplasia</i> — hyperplasie folliculaire (classe bénigne)
FL	<i>Follicular Lymphoma</i> — lymphome folliculaire (classe maligne de la Phase 2)
GPU	<i>Graphics Processing Unit</i> — processeur graphique
H&E	Hématoxyline-éosine — coloration standard des lames
IUCT	Institut Universitaire du Cancer de Toulouse
LDGCB	Lymphome Diffus à Grandes Cellules B
LLC	Leucémie Lymphoïde Chronique
MIL	<i>Multiple Instance Learning</i> — apprentissage à instances multiples
MLP	<i>Multi-Layer Perceptron</i> — perceptron multicouche
OCR	Oncopole Claudius Regaud
RAM	<i>Random Access Memory</i> — mémoire vive
ReLU	<i>Rectified Linear Unit</i> — fonction d'activation
ROC	<i>Receiver Operating Characteristic</i>
RSE	Responsabilité Sociétale et Environnementale
SAM	<i>Segment Anything Model</i> (SAM2, utilisé par AtlasPatch)
SET	Systèmes Embarqués et Télécommunications (spécialité d'ingénieur)
THQSE	Très Haute Qualité Sanitaire, Sociale et Environnementale (label)
ViT	<i>Vision Transformer</i>
VRAM	<i>Video Random Access Memory</i> — mémoire de la carte graphique
WSI	<i>Whole Slide Image</i> — lame virtuelle
WSL	<i>Windows Subsystem for Linux</i>

Glossaire

Accuracy Proportion de lames correctement classées par le modèle, pour un seuil de décision fixé (ici 0,5).

Arrêt anticipé (early stopping) Technique qui stoppe l'entraînement du modèle lorsque la performance sur le dataset de validation ne s'améliore pas pendant un nombre fixé d'épochs (patience), et conserve la meilleure version du modèle pour le dataset de test.

AtlasPatch Outil open-source développé pour les whole slide images (WSI) afin de les découper en patchs et de les encoder par un modèle de fondation.

AUC (Area Under the Curve) Mesure de la capacité du modèle à donner un score plus élevé à une lame maligne qu'à une lame bénigne, indépendamment du seuil de décision.

Biopsie Prélèvement d'un échantillon de petite taille afin de l'observer au microscope.

Descente de gradient Algorithme d'optimisation qui ajuste les poids des neurones dans la direction qui diminue le plus l'erreur mesurée par la perte.

Dropout Technique de machine learning qui consiste à éteindre aléatoirement une partie des neurones pour que le modèle ne s'appuie pas sur quelques neurones seulement, ce qui limite le surapprentissage.

Encodeur Voir la définition de Modèle de fondation.

Entropie croisée Fonction de perte pénalisant l'écart entre la probabilité prédite et la vraie classe ; sensible à la confiance des prédictions, elle pénalise très fortement une erreur commise avec beaucoup de confiance.

Epoch Un passage complet du modèle sur l'ensemble des données d'entraînement.

Feature Composante du vecteur produit par l'encodeur pour décrire le contenu d'un patch.

Fuite de données Passage d'informations entre l'ensemble d'entraînement et l'ensemble de test, qui gonfle artificiellement les performances mesurées.

Hyperplasie folliculaire Réaction inflammatoire bénigne du ganglion, dont l'aspect est similaire au lymphome folliculaire.

Immunohistochimie Technique de laboratoire permettant de mettre en évidence, à l'aide d'anticorps, des protéines spécifiques pour préciser un diagnostic.

Lymphome folliculaire Cancer du tissu lymphoïde, cible de la Phase 2 du stage.

Lymphome de Hodgkin Type de cancer lymphoïde, plus facilement détectable que le lymphome folliculaire.

MIL (Multiple Instance Learning) Cadre d'apprentissage où une lame est traitée comme un ensemble (sac) de patchs et classée globalement plutôt que patch par patch. Chaque patch se voit attribuer un poids qui représente sa contribution à la prédiction finale, ce qui permet au modèle de porter son attention sur les régions les plus informatives de la lame.

Modèle de fondation Grand réseau de neurones pré-entraîné sur d'immenses jeux de données (ici H-optimus-0, H-optimus-1 et Virchow2, spécialisés en pathologie) ; il produit un vecteur de features décrivant chaque patch.

Patch Petite image (ici 256×256 pixels) extraite d'une image numérique.

Perte (loss) Voir la définition d'Entropie croisée.

Pièce chirurgicale Prélèvement étendu issu du bloc opératoire, correspondant à l'extraction d'un ganglion entier ou d'un large fragment.

ReLU Fonction d'activation ne conservant que les valeurs positives, qui introduit de la non-linéarité et permet au réseau d'apprendre des relations complexes.

Seed Graine du générateur aléatoire ; la fixer rend les expériences reproductibles.

Softmax Fonction transformant un ensemble de scores en probabilités positives dont la somme vaut 1.

StratifiedGroupKFold Fonction de la bibliothèque Python scikit-learn qui découpe les données en préservant les proportions des classes tout en gardant toutes les lames d'un même patient dans un unique ensemble.

Surapprentissage (overfitting) Situation où le modèle mémorise des éléments non pertinents du jeu d'entraînement, ce qui dégrade ses performances sur de nouvelles données.

Taux d'apprentissage (learning rate) Taille du pas effectué à chaque mise à jour lors de la descente de gradient.

Vision Transformer (ViT) Architecture de réseau de neurones de type *transformer* appliquée aux images, sur laquelle reposent H-optimus et Virchow2.

Vote majoritaire Décision au niveau de la lame obtenue en comptant les patchs classés bénins et malins, la classe majoritaire l'emportant.

WSI (Whole Slide Image) Lame virtuelle (lame numérisée), stockée sous forme de pyramide multi-résolution (plusieurs grossissements).

Résumés

Résumé (français)

Ce travail porte sur la classification de **lames numériques** par intelligence artificielle afin de distinguer automatiquement une **hyperplasie folliculaire** (bénigne) d'un **lymphome folliculaire** (malin). Le travail s'est déroulé en deux temps. La méthode a d'abord été construite sur un premier jeu de données (lymphome de Hodgkin), puis appliquée au jeu cible (lymphome folliculaire), seules les données changeant.

Le pipeline commence par l'extraction de patchs (AtlasPatch), puis l'encodage par **modèle de fondation** (H-optimus, Virchow2), un découpage des données par patient, et enfin une classification par **apprentissage à instances multiples (MIL)**. Dans la première phase, les performances augmentent considérablement du vote majoritaire avec ResNet50 (accuracy 0,746) au MIL à attention avec H-optimus-1 (0,932), ce qui permet de valider l'approche. En seconde phase, sur le lymphome folliculaire, j'ai développé un modèle à double encodeur, qui regarde le type de lame et utilise le modèle le plus performant en fonction de ce type : les biopsies vers un modèle 20× et les pièces chirurgicales vers un modèle multi-échelle. Ce modèle atteint la meilleure accuracy (0,962) de tout mon stage et une AUC de 0,989, avec seulement cinq erreurs sur les 130 lames. En plus de ces résultats très satisfaisants, le pipeline est réutilisable et transposable à un autre problème sans modifier sa chaîne de traitement.

Mots-clés : MIL, lymphome folliculaire, hyperplasie folliculaire, modèle de fondation, lames numériques.

Abstract (English)

This work concerns the AI-based classification of **digital slides** and enables the automatic distinction between **follicular hyperplasia** (benign) and **follicular lymphoma** (malignant). The process was carried out in two stages. The method was first developed on an initial dataset (Hodgkin lymphoma), then applied to the target dataset (follicular lymphoma), with only the data changing.

The pipeline begins with patch extraction (AtlasPatch), followed by encoding through a **foundation model** (H-optimus, Virchow2), a patient-level split, and finally classification via **multiple instance learning (MIL)**. In the first stage, performance increases considerably from majority voting with ResNet50 (accuracy 0.746) to attention-based MIL with H-optimus-1 (0.932), which validates the approach. In the second stage, on follicular lymphoma, a dual-encoder model examines the slide type and applies the best-performing model for each type : biopsies are routed to a 20× model and surgical specimens to a multi-scale model. This model achieves the best accuracy (0.962) of my entire internship and an AUC of 0.989, with only five errors across the 130 slides. In addition to these highly satisfactory results, the pipeline is reusable and transferable to a different problem without changing its processing chain.

Keywords : MIL, follicular lymphoma, follicular hyperplasia, foundation model, digital slides.