Réduction du biais lié à la coloration des coupes histologiques de cancer du sein pour le développement d'outils de classification par intelligence artificielle

Camille Franchet

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Réduction du biais lié à la coloration des coupes histologiques de cancer du sein pour le développement d'outils de classification par intelligence artificielle

---

Thèse présentée et soutenue, le 9 décembre 2024 par

Camille FRANCHET

École doctorale
BSB - Biologie, Santé, Biotechnologies

Spécialité
Cancérologie

Unité de recherche
CRCT - Centre de Recherches en Cancérologie de Toulouse

Thèse dirigée par
Pierre BROUSSET

Composition du jury
M. Julien CALDERARO, Président, Université Paris Est Créteil
M. Thomas WALTER, Rapporteur, Mines Paris - PSL
Mme Sandrine MOUYSSET, Examinatrice, Université Toulouse III - Paul Sabatier
M. Emmanuel BACRY, Examinateur, CNRS Paris-Centre
Mme Magali LACROIX-TRIKI, Examinatrice, Institut Gustave Roussy
M. Pierre BROUSSET, Directeur de thèse, Université Toulouse III - Paul Sabatier

Remerciements

Au Pr Thomas Walter,

Je tiens à exprimer ma gratitude au Professeur Thomas Walter, rapporteur de cette thèse, pour le temps précieux consacré à l'évaluation de mon travail. Vous m'avez fait l'honneur d'accepter de juger ce travail qui, je l'espère, était à la hauteur de votre expertise. Merci d'avoir pris le temps d'évaluer minutieusement cette thèse et d'avoir partagé votre expertise et votre culture dans le domaine des biais associés au machine learning.

Au Pr Julien Calderaro,

Je remercie chaleureusement le Professeur Julien Calderaro, rapporteur de cette thèse, pour son engagement et ses précieux conseils. Vous m'avez fait l'honneur d'étudier avec attention mon travail de thèse, à l'aune de votre expérience dans le développement et la validation de modèles d'IA en anatomie pathologique. Je vous suis infiniment reconnaissant.

Aux membres du jury,

Je tiens à exprimer ma plus sincère reconnaissance aux membres du jury, les Professeurs Sandrine Mouysset et Emmanuel Bacry, ainsi que le Docteur Magali Lacroix-Triki, pour l'honneur qu'ils m'ont fait en évaluant mon travail de thèse. C'est un privilège pour moi d'avoir pu présenter mes recherches devant un jury d'une telle qualité. A mon maître ès sénopathologie, Magali, merci pour tout depuis le début. Ta confiance sans faille est une source de motivation intarissable.

Au Pr Pierre Brousset,

Je vous remercie pour la confiance que vous m'avez accordée, dès le lendemain de ma thèse de médecine, en acceptant de diriger ce travail. Je vous suis extrêmement reconnaissant d'avoir su recruter très tôt, au sein du service, des data scientists auxquels je dois l'essentiel de ma culture dans le domaine de l'IA. Vous avez investi en conséquence dans une infrastructure informatique de qualité qui a permis de mener à bien nos premiers projets d'IA. Certes, j'aurais mis plus de temps que prévu à finaliser ce travail, mais votre indéfectible soutien m'a permis de rebondir quand le projet initial s'est finalement transformé en une exploration inattendue du domaine vaste et complexe des biais en IA.

À Robin,

Mon alter ego data scientist, collègue et ami. A la suite d'Arnaud, tu m'as transmis toutes tes

connaissances, ainsi que tes doutes sur le deep learning, le machine learning et tellement d'autres domaines. Ce travail est aussi le tien. Co-premier auteur de l'article, je te remercie infiniment pour tous tes efforts face à mes demandes incessantes de visualisation des résultats. C'est un réel plaisir de travailler avec toi. Maintenant que tu es inscrit en thèse de sciences, j'espère qu'une géniale architecture appelée Robin-Net va enfin voir le jour.

À FX, Gabrielle, Stéphanie et Rodica,

Un très grand merci groupé à vous quatre. On l'a un peu oublié mais, Gabrielle, tu as été impliquée très tôt dans ce travail quand nous avons dû démonter et remonter l'intégralité des lames de la PACS04, une galère. Stéphanie et Rodica, quelle énergie vous avez déployée dans le projet APRIORICS ! Cela force le respect. Je vous dois toute la production de données de ce projet pharaonique que j'espère très fructueux à court terme. FX, tu es présent depuis le début, depuis les premiers travaux avec Arnaud jusqu'au projet APRIORICS et à tous les projets qui débutent.

À tous mes collègues,

Dont un bon nombre sont mes amis. Merci à tous. C'est un plaisir de travailler à vos côtés, dans la joie et la bonne humeur. J'inclus ici les collègues chercheurs de l'IRIT et du CRCT. A mon « agneaulesque » et brillantissime élève, Salvatore. J'espère pouvoir continuer à t'enseigner la pathologie mammaire pendant encore longtemps. A Stéphane, c'est un plaisir de faire partie de ton encadrement de thèse. Bientôt la fin de la review !

À Céline,

Ma femme d'amour qui me soutient dans toutes les situations. Je te dois notre famille, et cette stabilité qui nous permet de construire sereinement l'avenir ensemble. Je suis très heureux et fier de partager ma vie avec toi.

À mes enfants,

Mes petits choux, Victor, Haydée et Diane. Vous me rendez fier chaque jour qui passe. Vous êtes mon indéfectible îlot de bonheur. Je vous aime très fort.

À ma famille,

Mon éternelle source d'inspiration. Famille percheronne ou tourangelle, quel bonheur de ressentir votre soutien et votre confiance au quotidien. Merci pour tout.

Réduction du biais lié à la coloration des coupes histologiques de cancer du sein pour le développement d'outils de classification par intelligence artificielle

TABLE DES MATIERES

I. INTRODUCTION

A. CONTEXTE

L'émergence de la pathologie digitale est en train de bouleverser de manière majeure la pratique de l'anatomopathologie en France et dans le monde. En effet, l'apparition de scanners de lames il y a une trentaine d'années a permis de passer d'une visualisation de l'image microscopique sur un support physique - la coupe fine de tissu déposée sur une lame de verre puis colorée - à l'aide d'un microscope, à une visualisation de l'image numérique à très forte résolution sur un écran. Outre la modification du processus de visualisation, c'est surtout les opportunités qu'offre l'image numérique en termes de flux, d'archivage, d'annotation et d'analyse qui ouvre des perspectives enthousiasmantes. L'essor de la numérisation des lames dans les laboratoires de pathologie s'est fait de concert avec celui des techniques d'analyse d'image et de vision par ordinateur telles que les réseaux neurones convolutifs au cours des années 2010.

C'est dans ce contexte qu'a débuté mon travail de thèse de sciences, dès 2016, qui avait alors pour principal objectif d'affiner l'évaluation pronostique d'une catégorie de cancer du sein afin de mieux en guider le traitement, tout en identifiant des patterns morphologiques associés au pronostic. A l'époque, la notion de « Tissue Phenomic » connaissait un certain succès, portée par Gerd Binnig, prix Nobel de physique 1986 et fondateur de l'entreprise Definiens AG. Les outils développés par cette entreprise proposaient une approche « orientée objet » de la classification d'images microscopiques, permettant notamment de découvrir des traits morphologiques interprétables associés à un phénotype [1]. Les découvertes associées à ce concept et à l'utilisation de cette suite logicielle semblaient prometteuses et furent une source d'inspiration au début de mon travail [2,3]). Cependant, la segmentation des « objets » dont l'outil extrayait des caractéristiques était relativement mauvaise, rendant la pertinence biologique et la validité externe des résultats difficile à établir. En parallèle, l'accessibilité croissante des méthodes de deep learning, notamment grâce à la mise à disposition des frameworks TensorFlow et PyTorch, a rendu caduques la segmentation et la classification d'image par des méthodes conventionnelles de traitement d'image. J'ai eu la chance, à ce moment-là, de me trouver dans un laboratoire qui a investi tôt dans de bonnes cartes graphiques et dans le recrutement de stagiaires et thésards issus d'écoles d'ingénieurs. C'est au contact de ces personnes (Arnaud Abreu en premier lieu, Thomas Fel, puis Robin Schwob) que j'ai pu découvrir le machine learning, deep learning compris, et construire ma culture mathématique, critique et « philosophique » du domaine. C'est finalement en avançant dans mon travail de thèse avec un regard critique et une obsession de la visualisation des résultats que j'ai découvert un pan entier

et peu médiatisé de l'intelligence artificielle : les biais et les méthodes d'atténuation de ces biais. Autrement dit, les pièges qui peuvent donner l'impression, en toute bonne foi, qu'un modèle est satisfaisant alors qu'il ne l'est pas. La détection et la compréhension de ces biais nécessitent souvent une connaissance conjointe du métier de pathologiste et du fonctionnement réel des algorithmes de machine learning. C'est à cette interface que je me suis épanoui.

Ce travail a commencé avec l'hypothèse qu'une approche d'analyse d'image poussée pourrait identifier des caractères morphologiques associés au pronostic dans les images microscopiques de cancer du sein. Ce faisant, elle pourrait outrepasser les capacités actuelles d'évaluation du pronostic et devenir partie intégrante de la prise de décision thérapeutique en sénologie. Même si l'objectif du projet a évolué au cours de mon travail, il reste néanmoins très associé au domaine de la pathologie, au cancer du sein à de l'essor de l'intelligence artificielle dans ces domaines.

B. TECHNIQUE EN ANATOMIE PATHOLOGIQUE

Compte tenu de l'impact de la technique, et notamment de toute la phase pré-analytique, sur l'aspect final des images microscopiques, il semble ici pertinent de décrire brièvement les différentes étapes de prise en charge d'un prélèvement jusqu'à la création d'une lame colorée.

B.1 DUREE D'ISCHEMIE FROIDE

Il s'agit de la durée écoulée entre la dévascularisation d'un prélèvement et sa mise en contact avec un fixateur. Durant cette période, s'enclenche un processus de dégradation tissulaire lié à l'arrêt de l'apport d'oxygène aux cellules. Si cette période dure trop longtemps, on pourra observer des phénomènes de putréfaction.

Afin d'éviter cela, la durée d'ischémie froide doit être minimale. Pour les prélèvements biopsiques, on considère que la durée d'ischémie froide ne doit pas excéder 10 minutes. Pour les pièces opératoires, la durée d'ischémie froide ne doit pas excéder 1 heure. En cas d'impossibilité à prendre en charge un prélèvement chirurgical dans un délai raisonnable, la mise sous vide ou la conservation au réfrigérateur à 4°C diminuent drastiquement la dégradation du tissu [4].

A noter que pour les pièces opératoires volumineuses, une ouverture de la pièce est nécessaire afin de faciliter et homogénéiser la fixation et d'exposer la surface tumorale afin qu'elle soit directement au contact du fixateur. Cette étape de préparation à la fixation est indispensable à la qualité des images microscopiques finales et, par conséquent, à la recherche de biomarqueurs dans des conditions optimales.

B.2 FIXATION AU FORMOL

Elle a pour but de bloquer les enzymes endogènes et la pullulation microbienne à l'origine des phénomènes de putréfaction. Elle doit aussi respecter les structures cellulaires et tissulaires dans un état proche de l'état physiologique, tout en préservant leur réactivité pour les études immunohistochimiques. Enfin, elle doit préparer à l'inclusion en paraffine [5].

Le fixateur recommandé et utilisé de manière quasiment systématique à l'échelle internationale est le formol. Idéalement une solution de formaline à 10% (équivalent à une solution de formaldéhyde à 4%), neutre, tamponnée. Le volume du fixateur doit être au moins de 10 fois le volume de prélèvement et le récipient doit être suffisamment grand pour éviter les déformations des pièces opératoires. Comme indiqué précédemment, une étape de préparation à la fixation est indispensable pour certains prélèvements : ouverture des organes creux, tranchage des organes pleins volumineux.

Le formol crée des ponts méthylène entre les molécules entre les lésions peptidiques. Pour obtenir un bon compromis entre la pénétration du formol dans les tissus et sa fixation, il convient de maintenir le formol à température ambiante pour que la réaction chimique se fasse de manière optimale. La durée de fixation est de 8 à 24h pour les biopsies et 24 à 48h pour les pièces opératoires. Une fois fixées, les pièces opératoires sont examinées et échantillonnées au cours de l'étape de macroscopie.

B.3 MACROSCOPIE DES PRELEVEMENTS FIXES

Il s'agit d'une observation soigneuse des prélèvements permettant une description précise des lésions observées et guidant l'échantillonnage ciblé du prélèvement (découpe). Les échantillons prélevés doivent être représentatifs des lésions et permettre de répondre à toutes les questions nécessaires à la bonne prise en charge du patient. Chaque échantillon est positionné dans un réceptacle ajouré appelé « cassette » qui mesure environ 3 x 2 x 0,4 cm (Figure 1). Cette cassette est assez grande pour y positionner une tranche de quelques millimètres d'épaisseur et sa taille est compatible avec une lame de verre. Chaque cassette est annotée avec le numéro de dossier du patient et un numéro unique. Pour chaque numéro, le contenu de la cassette est consigné.

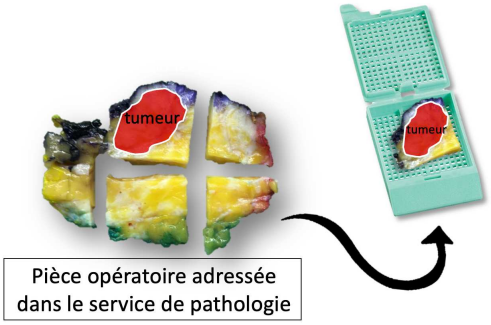


Figure 1 : Macroscopie : échantillon tumoral coupé puis positionné dans une cassette.

B.4 IMPREGNATION ET INCLUSION EN PARAFFINE

Même fixés dans le formol, les tissus restent trop malléables pour que des coupes très fines puissent être réalisées. L'imprégnation consiste à faire durcir un tissu en remplaçant l'eau qu'il contient par une matière rigide (à température ambiante), totalement hydrophobe et chimiquement inactive : la paraffine. Pour ce faire, les prélèvements sont déshydratés à l'éthanol (l'eau du prélèvement est progressivement remplacée par de l'alcool pur), lui-même substitué par un solvant de la paraffine appelé xylène. Enfin, l'étape d'imprégnation proprement dite consiste à remplacer le xylène par de la paraffine liquide à 60°C. A l'issue de cette étape, le contenu des cassettes est inclus dans un moule de taille adapté, puis rempli de paraffine liquide surmontée de la cassette, qui ne joue plus alors le rôle de réceptacle mais de support du prélèvement : c'est le bloc de paraffine (FFPE, pour *formalin-fixed paraffin-embedded*) (Figure 2).

Figure 2 : Bloc de paraffine contenant un fragment tissulaire fixé en formol (FFPE)

Sous cette forme, l'échantillon peut être conservé à température ambiante pendant des années sans risquer de se dégrader. Le bloc de paraffine est à la fois un moyen de conservation et le point de départ de toutes les techniques morphologiques, immunohistochimiques et moléculaires utilisées en pathologie quotidienne.

B.5 COUPE AU MICROTOME ET CREATION DE LAMES BLANCHES

De par sa dureté à température ambiante, le bloc FFPE peut être coupé en tranches de quelques micromètres d'épaisseur à l'aide d'un microtome, que l'on applique ensuite sur des lames de verre pour créer des lames blanches (Figure 3). Une lame blanche consiste en une coupe fine de tissu quasiment incolore enchâssée dans de la paraffine et posée sur une lame de verre. Après déparaffinage, elle peut être colorée par différents moyens : coloration standard à l'hématoxyline-éosine (HE), coloration spéciale, immunohistochimie. Pour chaque échantillon mis en cassette, une lame en coloration standard sera réalisée. Cette lame HE est le premier support physique de l'information morphologique. Son faible coût et son accessibilité expliquent son utilisation universelle dans tous les laboratoires de pathologie du monde.

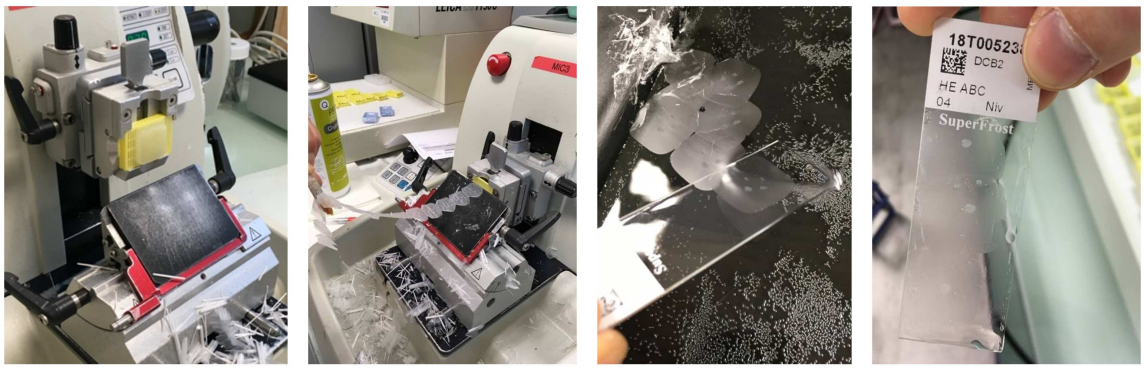


Figure 3 : Coupe d'un bloc FFPE au microtome et réalisation d'une lame blanche

B.6 LA COLORATION HEMATOXYLINE EOSINE (HE)

Nous allons aborder plus en détail la coloration HE du fait de son caractère central dans ce travail de thèse.

B.6.1 HISTOIRE DE LA COLORATION HE

En routine, dans un laboratoire de pathologie, tous les prélèvements tissulaires FFPE font l'objet d'une coloration HE. Cette coloration a été décrite pour la première fois en 1876 et s'est imposée comme la coloration standard. De manière extrêmement simplifiée, elle consiste en une coloration bleu violacé (hématoxyline) des noyaux et une coloration rose des cytoplasmes et du collagène (éosine). Pour apprécier toutes les nuances qui peuvent être rendues par cette simple

coloration, voir le très bel article de John K C Chan qui expose toute la beauté de la coloration HE en histologie [6].

L'hématoxyline est un composé naturel extrait du duramen de l'arbre de Campêche (*Logwood tree* en anglais - *Hematoxylon Campechianum* en latin) qui pousse naturellement en Amérique Centrale. Même si la synthèse d'hématoxyline est possible, elle ne l'est pas en quantité commerciale. De ce fait, toute l'hématoxyline utilisée de par le monde est d'origine naturelle. Le terme d'hématoxyline est largement employé, même si stricto sensu ce n'est pas l'hématoxyline qui est utilisée pour colorer les tissus mais sa forme oxydée, l'hématéine. L'hématéine peut se comporter comme un acide ou comme une base (il s'agit d'une molécule amphotérique). Elle présente une coloration rouge à pH acide et bleue à pH basique [7]. Pour colorer les noyaux, l'hématéine nécessite d'être associée à un mordant, un alun, afin de faciliter la coloration, d'où le terme d'hémalun [8]. Dans notre laboratoire, nous utilisons l'hématoxyline de Mayer modifiée selon Gill III (une formulation de l'hématoxyline parmi d'autres) qui utilise l'iodate de sodium comme méthode d'oxydation chimique et le sulfate d'aluminium comme mordant. Il s'agit d'une coloration progressive au cours de laquelle on contrôle la durée d'exposition au colorant afin d'ajuster le niveau de coloration souhaité. A noter qu'il existe également des méthodes régressives au cours desquelles on surcolore l'échantillon, avant de réaliser une étape de lavage. Enfin, la "différenciation" à l'aide d'une solution d'alcool et d'acide chlorhydrique permet de supprimer les marquages non spécifiques, et le passage dans l'eau ammoniaquée permet de stabiliser la coloration bleue.

L'éosine a été inventée par le chimiste allemand Heinrich Caro. C'est un dérivé de la fluorescéine qui tient son nom d'un surnom donné à Heinrich Caro (Eos) lorsqu'il était enfant [9]. Il s'agit d'un colorant anionique, dit également acide, qui a une affinité pour les éléments cellulaires chargés positivement (= cationiques ou basiques). Il colore le cytoplasme en rose et les autres éléments cellulaires basiques en rose/rouge plus ou moins vifs selon leur acidophilie. Sa formulation inclut souvent de l'acide acétique afin de réduire le pH et obtenir une nette charge positive des protéines. Différents types d'éosine existent. En anatomie pathologique, on utilise le plus souvent l'éosine Y qui colore les tissus de manière plus intense que l'éosine B.

B.6.2 VARIANTES (SAFRAN, ORANGE G, PHLOXINE)

Si la coloration HE constitue la référence pour l'examen histologique à travers le monde, les "recettes" et variantes sont nombreuses, notamment par ajout de colorants supplémentaires comme le safran (HES) ou l'Orange G qui vont apporter une teinte jaune-orangée au collagène. D'autre part, dans certains laboratoires, l'intensité du marquage à l'éosine est augmentée par ajout de phloxine, on parle alors de coloration HPS. Tous ces éléments mettent en évidence la

nature très artisanale de la coloration HE. Ceci permet de comprendre aisément la variabilité de coloration qui peut exister entre laboratoires, ou même dans le temps au sein d'un même laboratoire.

B.6.3 REALISATION DE LA COLORATION

Cette coloration est bien entendue automatisée dans notre laboratoire. Le protocole utilisé est le suivant :

- Déparaffiner et réhydrater :
    - *Déparaffinage* : Xylène 5 mn.
    - *Déshydratation :* Alcool absolu 5 mn.
    - *Réhydratation* : Eau courante pendant 5 à 10 mn.
- Hémalun de Mayer pur 5 minutes
- Eau
- Alcool chlorhydrique 0.3% pendant 15 secondes
- Eau
- Eau ammoniaquée à 0.4% pendant 15 secondes
- Eau
- Eosine-Orange G pendant 2,30 min
- Alcool
- Xylène

A l'issue de ce processus technique qui a permis de passer d'un prélèvement biopsique ou chirurgical à une ou plusieurs lames en coloration standard, l'examen microscopique peut débuter. Classiquement, on lit les lames à l'aide d'un microscope à fond clair (ou à lumière transmise). Une source de lumière traverse la lame de verre et la coupe colorée et l'image microscopique est alors observable à plusieurs grossissements via les oculaires.

C. LA PATHOLOGIE DIGITALE

C.1 CONTEXTE

L'apparition progressive, dans les trente dernières années, d'outils de numérisation des lames de verre, appelés scanners de lames, a permis de passer d'une visualisation de l'image microscopique sur un support physique - la coupe colorée sur une lame de verre - à l'aide d'un microscope, à une visualisation de l'image numérique à très forte résolution sur un écran [10,11]. Outre la modification du processus de visualisation, c'est surtout les opportunités

qu'offre l'image numérique en termes de flux, d'archivage, d'annotation et d'analyse qui ouvre des perspectives enthousiasmantes. L'essor de la numérisation des lames dans les laboratoires de pathologie s'est fait de concert avec celui des techniques d'analyse d'image et de vision par ordinateur telles que les réseaux neurones convolutifs au cours des années 2010.

C.2 FORMAT DES LAMES VIRTUELLES (WHOLE SLIDE IMAGES - WSI)

Les fichiers de lames virtuelles restent très volumineux à l'heure actuelle, les principaux avantages de la pathologie digitale sont la facilité de transmission d'images et de données pathologiques à travers le monde, l'archivage sûr, sécurisé et rapidement requêtable de spécimens histologiques et le développement d'outils d'analyse d'image qualitative et quantitative utilisables en pratique courante.

C.3 FORMATS DE FICHIERS D'IMAGES NUMERISEES

Une lame virtuelle, ou WSI, peut avoir une taille de plusieurs millions de pixels, ce qui pose d'importants problèmes d'archivage et de manipulation d'image. Le fichier est souvent si volumineux (plusieurs giga-octets) qu'il ne peut être contenu tout entier dans la mémoire vive d'un ordinateur, empêchant d'emblée son ouverture et sa modification. Les fabricants de scanners de lames utilisent donc un système de streaming naviguant au sein de ce que l'on appelle une image pyramidale (Figure 4). Ce système de pyramide où les couches à fort grossissement sont divisées en tuiles s'apparente à l'application Google Earth. Le logiciel de visualisation ne charge que les tuiles que l'utilisateur souhaite visualiser, ce qui permet de réduire énormément le volume de données à afficher. Des librairies python comme OpenSlide[1] ou cuCIM[2] permettent de prendre en charge ce type d'images.

A l'heure actuelle, un certain nombre de laboratoires de pathologie numérise leur activité de routine et visualise les cas *via* un système de gestion d'images. Plusieurs études ont montré que la lecture de lames virtuelles n'était pas statistiquement inférieure en terme de performance diagnostique, que la lecture de lames de verre au microscope [12–15]. Cependant, une revue systématique de la littérature a montré que la notion de surdiagnostic (à distinguer des faux positifs) était sous-représentée dans ces études [16–18]. Une étude abordant cette notion de manière indirecte montrait un taux de surdiagnostic de 3 % associé à la lecture sur WSI .[19] Dans le domaine de la sénologie, l'équipe de BJ Williams rapporte également d'excellents niveaux de concordance entre la lecture de biopsies et de pièces opératoires de sénologie sur WSI et sur lames de verre. Cependant, ils relèvent que la majorité des discordances provient du compte mitotique (plus difficile sur WSI) et de la détection de weddellites (microcalcifications

[1] https://openslide.org/api/python
[2] https://docs.rapids.ai/api/cucim/stable

d'oxalate de calcium dihydraté, translucides) [20]. Cette difficulté n'est pas rapportée dans une étude spécifiquement dédiée au compte mitotique réalisée par l'équipe de SA van Bergeijk, qui rapporte un R^2 de 0,85 et 0,83 sur biopsies et pièces opératoires, respectivement [21].

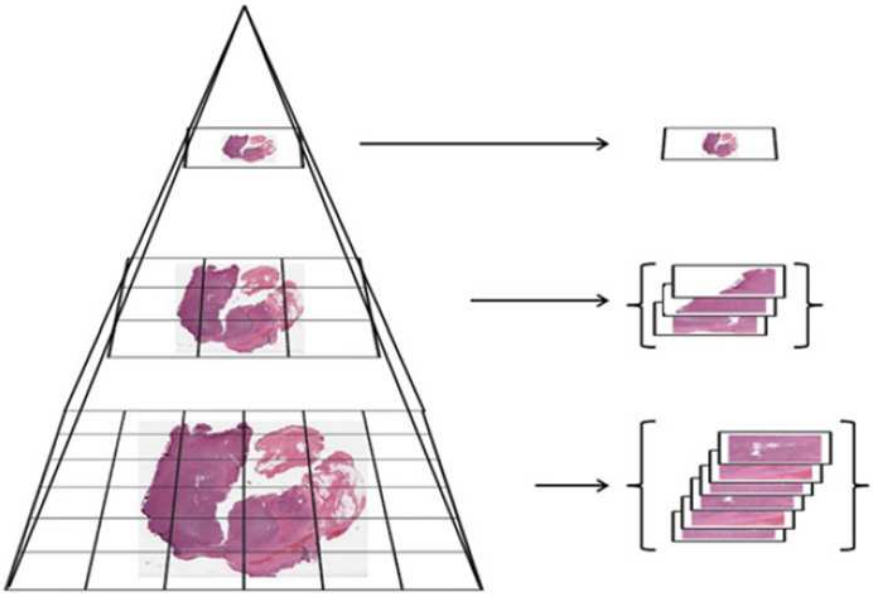


Figure 4 : Image pyramidale (WSI) : représentation d'un fichier de lame entière numérisée sous forme de trois couches à des grossissements différents. Chaque couche est enregistrée sous forme d'une image individuelle ou d'un ensemble d'images (tuiles) (adapté de [22]).

C.4 DEVELOPPEMENT D'OUTILS D'INTELLIGENCE ARTIFICIELLE EN PATHOLOGIE

Outre la facilité de partage et d'accès aux images microscopiques qu'offre la pathologie digitale, la numérisation des images microscopiques est également un prérequis indispensable au développement d'outils d'analyse d'image automatisée. De surcroit, l'essor de la pathologie digitale s'est fait de concert avec l'augmentation de la capacité de calcul des ordinateurs et le développement des approches d'analyse d'image par *deep learning*. Le deep learning est une branche du *machine learning* qui utilise des réseaux de neurones (artificiels) à plusieurs couches. Le machine learning, ou apprentissage automatique, regroupe toutes les techniques visant à donner aux machines la capacité d'« apprendre » à partir de données, *via* des modèles mathématiques[3]. Ces approches de machine learning sont particulièrement intéressantes dans le domaine de la pathologie digitale compte tenu de la complexité des images et de leur contenu

[3] https://www.cnil.fr/fr/definition/apprentissage-automatique

biologique. Cette complexité et cette variabilité sont très difficiles à appréhender par des approches de traitement d'image conventionnelles.

Il existe de nombreux modèles différents de machine learning et de deep learning qui répondent à des questions différentes : classification d'image, détection d'objet, segmentation d'objet, clustering. Ces modèles reposent sur deux grands types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé est très largement utilisé en médecine. Il consiste à utiliser des images labellisées ou annotées afin de « guider » l'algorithme. Le modèle d'IA apprend alors de chaque exemple, notamment en modifiant ses paramètres (les poids de chaque neurone s'il s'agit de deep learning) en cas d'erreur (Figure 5). L'apprentissage non supervisé correspond à des approches de clustering et reste assez peu utilisé en médecine, même si on pourrait discuter de son intérêt potentiel dans la détection des biais[4]. Au-delà de cette dichotomie simpliste, il existe dorénavant un ensemble de niveaux de supervision qui comble peu à peu l'écart entre les approches purement supervisées et purement non supervisées [23,24]. Les applications de ces modèles d'IA en pathologie sont nombreuses : assistance diagnostique, détection d'objet, évaluation de biomarqueurs, prédiction d'anomalies moléculaires, prédiction du pronostic, prédiction de la réponse au traitement, assurance qualité [24–26].

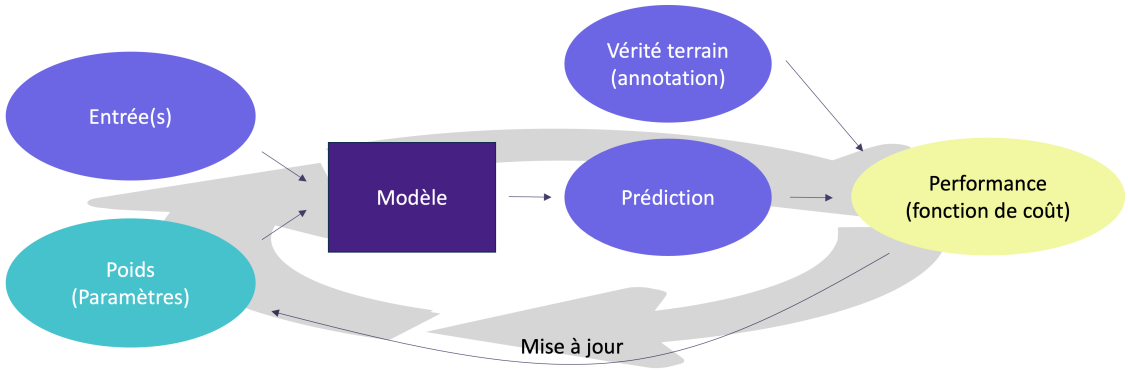


Figure 5 : Représentation schématique de l'apprentissage machine : un modèle produit une prédiction à partir d'une donnée d'entrée. Cette prédiction est comparée à la vérité terrain (annotation ou labellisation) et l'erreur est quantifiée grâce à la fonction de coût (loss). L'erreur est rétro-propagée afin d'ajuster les poids du modèle (paramètres). Ce cycle est répété durant toute la durée de l'apprentissage. *Source : cours SIDES – Bases du machine learning et des réseaux de neurones – C. Franchet & J. Calderaro.*

Dans notre travail, nous avons utilisé des réseaux de neurones convolutifs (CNN pour Convolutional Neural Network). Ces réseaux de neurones détectent des patterns morphologiques dans les images microscopiques *via* l'utilisation de filtres de convolution. Au cours de l'apprentissage, ces filtres sont modifiés et leurs poids respectifs ajustés pour

[4] https://oecd.ai/en/catalogue/tools/unsupervised-bias-detection-tool

distinguer au mieux, par exemple, différentes classes d'images s'il s'agit d'une tâche de classification, parmi les données d'apprentissage. Le modèle est évalué après chaque cycle d'apprentissage (appelé *epoch*) sur le dataset de validation. Le taux d'erreur du modèle est estimé grâce à une fonction qui peut porter différents noms : *loss function*, *fonction de coût*, *fonction d'erreur*, *fitness*... Quel que soit le nom qu'on lui donne, l'objectif de l'apprentissage est de minimiser le résultat de cette fonction, tant sur le dataset d'apprentissage que sur celui de validation.

Ce chapitre ne saurait couvrir toute la théorie sous-jacente à l'utilisation des réseaux de neurones en général, et des CNN en particulier. De nombreuses ressources pédagogiques accessibles et gratuites sont disponibles sur internet. J'indique ici les noms de deux chaines YouTube d'excellente qualité pour appréhender les concepts sous-jacents à l'apprentissage machine : StatQuest de Josh Starmer[5] et 3Blue1Brown de Grant Sanderson[6] ; ainsi que la formation fast.ai de Jeremy Howard[7] pour l'approche pratique.

Nous listerons simplement ici les hyperparamètres les plus importants pour l'étape de *fine tuning* des modèles de deep learning. Les hyperparamètres sont des paramètres utilisés pour configurer un modèle de machine learning. Ils servent à contrôler le processus d'apprentissage et sont définis par l'utilisateur. Ils ne doivent pas être confondus avec les paramètres du modèle, qui correspondent notamment aux poids de chaque neurone appris par le modèle au cours de la phase d'apprentissage.

- **Architecture du modèle** : inclut notamment le nombre de neurones et de couches du modèle ainsi que les types de connexion entre les couches. L'architecture du modèle à un lien direct avec sa complexité.
- **Fonction de loss** : c'est la fonction qui permet de quantifier l'erreur du modèle. Elle doit être désignée précisément car elle conditionne totalement la qualité de l'apprentissage.
- **Nombre d'epochs** : correspond au nombre de fois que le modèle rencontrera l'ensemble du dataset d'apprentissage au cours de l'apprentissage. Ce nombre d'epochs peut-être défini arbitrairement, ou bien être associé à une notion d'*early stopping* afin d'arrêter l'apprentissage quand le résultat de la fonction de loss sur le dataset de validation stagne ou augmente.
- **Taux d'apprentissage, ou *learning rate*** : représente la force avec laquelle un modèle prend en compte ses erreurs. Un learning rate faible diminue la vitesse d'apprentissage et ne permet pas d'explorer de manière optimale les possibilités du modèle. Un learning

[5] https://www.youtube.com/@statquest
[6] https://www.youtube.com/@3blue1brown
[7] https://course.fast.ai/

rate fort rend l'apprentissage anarchique et risque de ne jamais atteindre un taux d'erreur satisfaisant.

- **Programmation du learning rate** : représente l'évolution du learning rate au cours de l'apprentissage. Classiquement, on commence avec un learning rate fort (exploration des possibilités du modèle) puis on le diminue progressivement (se rapprocher du taux d'erreur minimal). Dans nos expériences, nous avons souvent utilisé un programme de learning rate basé sur un cycle d'augmentation, suivi d'une diminution progressive (one cycle learning rate) qui offre de bonne performance pour une durée d'apprentissage relativement courte.
- **Momentum** : il est très lié au learning rate. Il a pour but de définir l'ajustement de la force et de la direction de l'étape suivante en fonction de l'étape précédente.
- ***Batch normalization, dropout, weight decay*** : ce sont des éléments de régularisation. Ils améliorent les performances et la généralisation du modèle en diminuant le risque de surapprentissage.

Le réglage fin des hyperparamètres s'appelle le *fine tuning* du modèle. Il ne s'agit pas d'une science exacte mais plutôt d'un processus empirique, par essais/erreurs. L'identification des meilleurs hyperparamètres pour un modèle peut aussi se faire via des approches de type *grid search* qui permettent de tester de multiples combinaisons d'hyperparamètres. Pour l'entraînement de modèles complexes, c'est une étape très longue qui remplace difficilement l'expérience du *data scientist*.

D. GENERALITES SUR LES CANCERS DU SEIN

Cette introduction sur les cancers du sein se veut synthétique, permettant essentiellement de rappeler quelques notions générales et de bien saisir le rationnel du projet initial et son évolution au cours du travail de thèse.

D.1 ÉPIDEMIOLOGIE DES CANCERS DU SEIN

Les dernières données épidémiologiques concernant le cancer du sein publiées par l'Institut National du Cancer[8] font état de 61 214 nouveaux cas en France en 2023, confirmant une incidence en hausse constante depuis 1990. L'âge médian au moment du diagnostic était de 64 ans. On recensait 12 600 décès en 2021, avec une baisse de la mortalité de 1,3 % par an entre 2011 et 2021 (Figure 6).

8 https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein

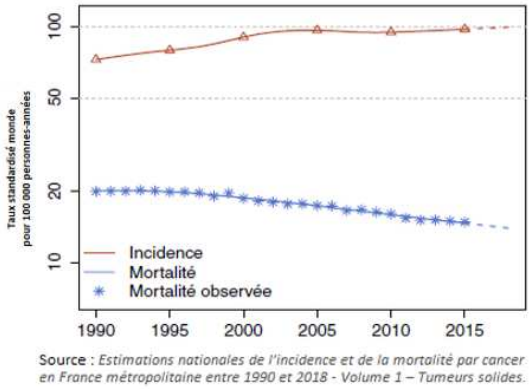


Figure 6 : Taux d'incidence et de mortalité par cancer du sein en France selon l'année (1990-2018)

Le cancer du sein reste au 1er rang des cancers incidents chez la femme, nettement devant les cancers recto-coliques et le cancer du poumon. C'est aussi la première cause de mortalité par cancer chez les femmes (18 % des décès par cancer chez les femmes en 2021) (Figure 7). Chez les hommes, le cancer du sein existe mais ne représente qu'environ 1% de l'ensemble des cancers du sein. Dans la suite de ce travail, le terme de « patientes », bien que linguistiquement inexact, sera utilisé pour refléter au mieux la pratique quotidienne de la prise en charge des cancers du sein.

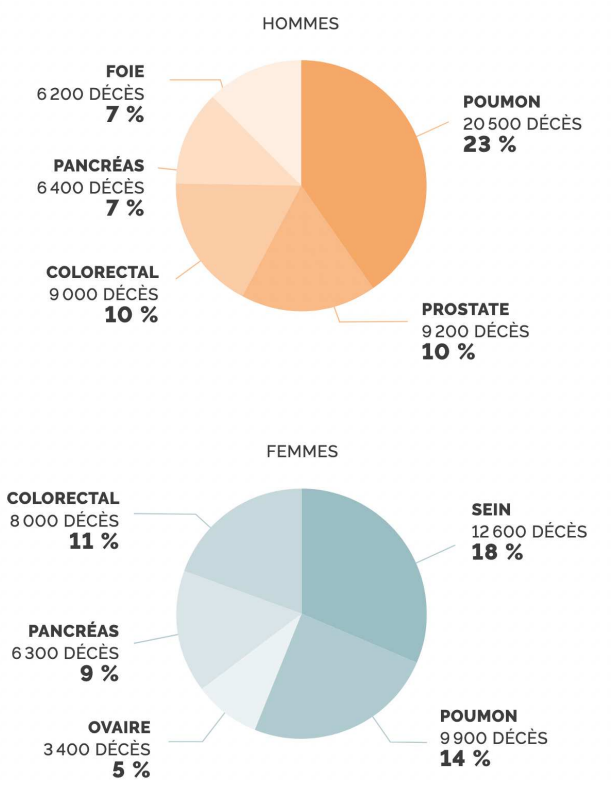


Figure 7 : Nombre de décès par cancer par sexe et par organe en 2021. Données de mortalité du Centre d'épidémiologie sur les causes médicales de décès (CépiDC). Infographie INCa 2024.

Les facteurs de risques de cancer du sein sont l'âge, la consommation d'alcool et de tabac, le surpoids et le manque d'activité physique, l'exposition aux hormones stéroïdes (naturelle ou liée à un traitement substitutif de la ménopause), l'absence d'allaitement et les prédispositions génétiques. A noter que l'alcool est le 1er facteur de risque exogène associé à l'apparition d'un cancer du sein[9].

D.2 DEPISTAGE ORGANISE DES CANCERS DU SEIN

Le programme de dépistage organisé des cancers du sein cible les femmes âgées de 50 à 74 ans sans symptôme ni facteur de risque particulier. Tous les 2 ans, elles sont invitées, par courrier, par mail ou par SMS envoyé par leur CPAM de rattachement, à réaliser une mammographie et un examen clinique des seins auprès d'un radiologue agréé. Le taux national de participation à ce dépistage est en baisse, avec un taux de 46,5 % pour la période 2022-2023. Ce taux faible serait en partie compensé par le fait qu'environ 11% de la population cible effectuerait des dépistages individuels[10].

D.3 DIAGNOSTIC DU CANCER DU SEIN

Devant toute suspicion clinique ou radiologique de cancer du sein, dans le cadre du dépistage organisé, d'un dépistage individuel ou par découverte d'une anomalie à l'autopalpation, un prélèvement biopsique est réalisé et examiné par un pathologiste afin de confirmer ou d'infirmer le diagnostic et de guider le traitement par l'évaluation des plusieurs biomarqueurs. Dès lors qu'un diagnostic de cancer du sein est posé, le cas est discuté en Réunion de Concertation Pluridisciplinaire (RCP), puis le diagnostic est annoncé à la patiente lors de la consultation d'annonce, en même temps que lui est faite une proposition thérapeutique, avec remise du Programme Personnalisé de Soin (PPS). Ce PPS donne à la patiente une vision globale de son parcours de soin, en incluant les soins de support. Cinq pour cent des patientes sont métastatiques au moment du diagnostic. Parmi les patientes présentant une maladie localisée au moment du diagnostic, environ un tiers va développer des métastases à distance malgré le traitement adjuvant [27].

D.4 GRANDES LIGNES DU TRAITEMENT DES CANCERS DU SEIN LOCALISES

Le traitement du cancer du sein localisé comprend deux axes principaux :

- Le traitement locorégional qui comprend la chirurgie et la radiothérapie. La chirurgie mammaire est associée à une chirurgie du creux axillaire homolatéral en cas de carcinome infiltrant. Le traitement par radiothérapie clôture le traitement locorégional. Elle est

[9] Panorama des cancers en France (2024) disponible sur www.ecancer.fr

[10] https://www.e-cancer.fr/Professionnels-de-sante/Depistage-et-detection-precoce/Depistage-du-cancer-du-sein/Le-programme-de-depistage-organise

obligatoire en cas de geste chirurgical conservateur. En incluant l'exérèse de la tumeur et l'éradication de potentielles cellules tumorales résiduelles présentes dans le sein et dans le creux axillaire, le traitement locorégional a pour objectif de réduire le risque de rechute locale ou locorégionale.

- Le traitement systémique qui inclut de manière variable l'hormonothérapie, la chimiothérapie, les thérapies ciblées et, dans certains cas l'immunothérapie. Il est dit « adjuvant » lorsqu'il est administré après la chirurgie, ou « néoadjuvant » lorsqu'il est administré avant. Le choix du traitement systémique est guidé par les caractéristiques cliniques de la patiente (comorbidités, antécédents) et, en grande partie, par les données de l'examen anatomo-pathologique de la biopsie et/ou de la pièce opératoire. La décision thérapeutique s'appuie sur les recommandations et référentiels régionaux, nationaux et internationaux émanant de groupes de travail ou de conférences de consensus tels que Saint Gallen ou Saint Paul de Vence [28,29].

Au stade précoce (localisé) de la maladie, l'examen anatomo-pathologique joue donc un rôle central dans le diagnostic du cancer, l'évaluation de son étendue (stade) et de son agressivité (grade histologique), l'évaluation de la qualité de l'exérèse chirurgicale et la quantification de biomarqueurs indispensable à la décision thérapeutique.

D.5 HISTOLOGIE DU SEIN ET CANCER

Histologiquement, le sein correspond à une glande apocrine hyperspécialisée associée à du tissu adipeux et collagénique en proportions variables. Le contingent glandulaire, en dehors de la période de lactation, représente environ 1 % de la masse du sein, tout le reste correspondant à du tissu adipeux, plus ou moins fibreux. Au niveau du mamelon s'abouchent les canaux galactophores de gros calibre. Telles les branches d'un arbre (l'arbre galactophorique), ces canaux se ramifient dans la profondeur du sein, jusqu'à aboutir, chez la femme, à l'unité terminale ductulo-lobulaire (UTDL). Au niveau des lobules, les acini sont bordés par une double assise cellulaire : une assise cellulaire épithéliale glandulaire interne, ou luminale, et une assise externe composée de cellules myoépithéliales qui associe des propriétés contractiles et une capacité de synthèse de la membrane basale. Cette membrane basale sépare la glande du tissu conjonctif environnant.

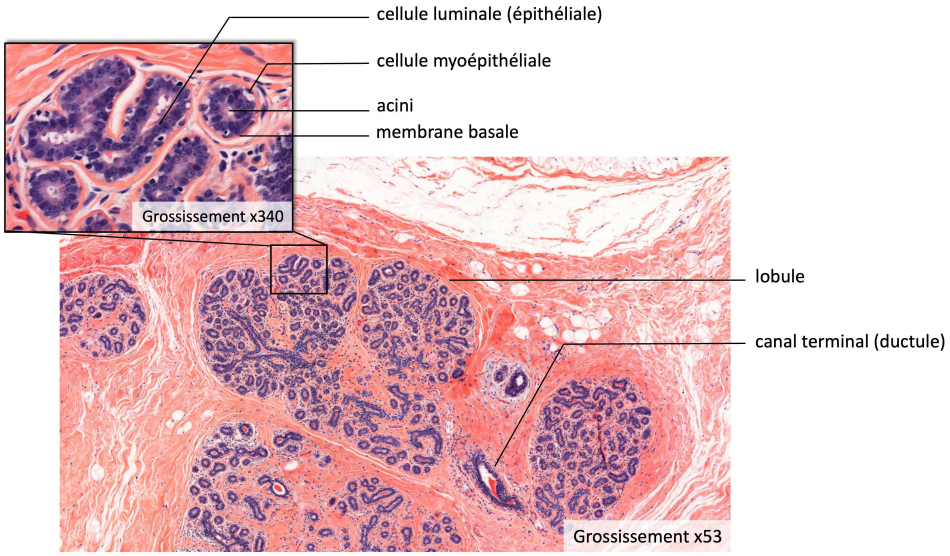


Figure 8 : Unité terminale ductulo-lobulaire

Une grande majorité des cancers du sein se développe à partir des cellules épithéliales glandulaires (cellules luminales). Lorsque la prolifération carcinomateuse se développe à l'intérieur des structures canalaires sans franchir la membrane basale, on parle de carcinome *in situ*. Lorsque la prolifération franchit la membrane basale et envahit le stroma sous-jacent, on parle de carcinome infiltrant (Figure 9). Dès lors, le cancer a un potentiel métastatique puisqu'il peut rejoindre la vascularisation lymphatique ou sanguine et disséminer au-delà du sein, vers les ganglions lymphatiques ou sous forme de métastases osseuses ou viscérales [30–32]. Ce sont les métastases à distance qui sont responsables du décès des patientes atteintes de cancer du sein.

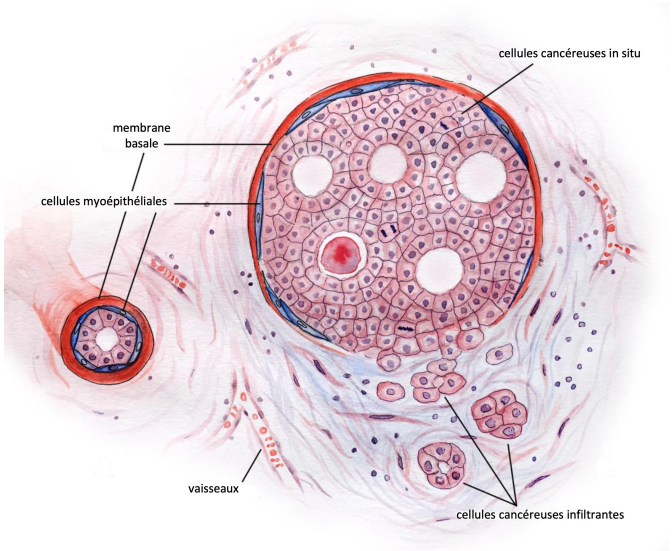


Figure 9 : Représentation schématique d'un acini et d'un foyer de carcinome canalaire in situ avec microinvasion (Adapté de [33])

D.6 DESCRIPTION HISTOPATHOLOGIQUE DES CARCINOMES INFILTRANTS DU SEIN

Le rôle du pathologiste ne consiste pas simplement à affirmer la présence de lésions carcinomateuses, in situ et/ou infiltrantes. Celui-ci doit également recueillir un certain nombre d'éléments qui permettront de décrire d'une part l'extension tumorale, d'autre part son agressivité biologique *via* l'identification de différents facteurs pronostiques et prédictifs. Mais avant tout, il est important de noter ici l'extraordinaire variété morphologique des cancers du sein, et ceci notamment dans une optique de développement d'outils de machine learning sur des images microscopiques de cancer du sein. Cette variété morphologique s'apprécie d'abord par l'évaluation du type histologique.

D.6.1 TYPE HISTOLOGIQUE

La dernière classification des cancers du sein par l'Organisation Mondiale de la Santé reconnait dix-huit types histologiques différents, auxquels s'ajoutent des variants (Tableau 1) [34].

Il n'est pas nécessaire ici de détailler l'ensemble des sous-types histologiques de cancer du sein, mais il semble pertinent de décrire au moins les deux premiers sous-types en fréquence et d'évoquer la très grande variabilité morphologique qui existe au sein même de ces sous-types.

Tableau 1 : Classification des types histologiques de carcinome infiltrant du sein [34]

Type histologique
Carcinome infiltrant de type non spécifique
Sous-types spéciaux
Carcinome lobulaire infiltrant
Carcinome tubuleux
Carcinome cribriforme
Carcinome mucineux
Cystadénocarcinome mucineux
Carcinome micropapillaire
Carcinome papillaire infiltrant
Carcinome avec différenciation apocrine
Carcinome métaplasique
Sous-types rares et de type glandes salivaires
Carcinome à cellules acineuses
Carcinome adénoïde kystique

Carcinome sécrétoire
Carcinome mucoépidermoïde
Carcinome polymorphe
Carcinome à cellules hautes avec polarité inversée
Carcinome papillaire infiltrant
Carcinome / tumeur neuroendocrine

Le type histologique de cancer du sein le plus fréquent est le carcinome infiltrant de type non spécifique. Il représente environ 80% des diagnostics de cancer du sein [34]. Quoique très fréquent, il s'agit en fait d'un diagnostic d'exclusion qui ne doit être retenu qu'après avoir évoqué et éliminé tous les autres types de cancers du sein. Par définition, ce type histologique renferme une grande variété d'aspects morphologiques et de nombreux variants sont décrits (médullaire, avec différenciation neuroendocrine, avec cellules géantes stromales ostéoclaste-like, pléomorphe, choriocarcinomateux, avec différenciation mélanique, oncocytaire, riche en lipides, riche en glycogène et sébacé). Ils sont également éminemment variables en termes de grade histologique, d'immunophénotype et de caractéristiques moléculaires.

Le second type histologique, et premier sous-type spécial en fréquence, est le carcinome lobulaire infiltrant. Il représente environ 15% des diagnostics de cancer du sein [34]. Il est classiquement constitué de cellules discohésives, de taille petite à moyenne, peu atypiques, et se disposant en file indienne. Au-delà de cette description du carcinome lobulaire infiltrant classique, il existe une grande diversité d'aspects morphologiques et plusieurs variants décrits : massif, alvéolaire, pléomorphe et tubulo-lobulaire.

Parmi les autres sous-types spéciaux, certains présentent des particularités morphologiques flagrantes comme la présence de flaques de mucus bleutées dans les carcinomes mucineux ou de zones tumorales d'apparence mésenchymateuse dans les carcinomes métaplasiques (Figure 10). Ces éléments méritent d'être connus car, bien que peu fréquents, peuvent être facilement identifiés par des outils d'analyse d'image. D'autres sous-types spéciaux sont de diagnostic plus délicat, comme les carcinomes tubuleux qui se distinguent peu des carcinomes infiltrants de type non spécifique les mieux différenciés (grade I), même si leur pronostic semble encore meilleur en analyse multivariée [35].

En termes de valeur pronostique, d'autres sous-types histologiques semblent associés de manière indépendante à un bon ou un mauvais pronostic, tels les carcinomes médullaires ou métaplasiques, respectivement [36]. Les autres sous-types histologiques peinent à montrer une valeur pronostique indépendante, même si certaines associations ressortent fréquemment en

analyse univariée, notamment celle entre carcinomes mucineux et bon pronostic.

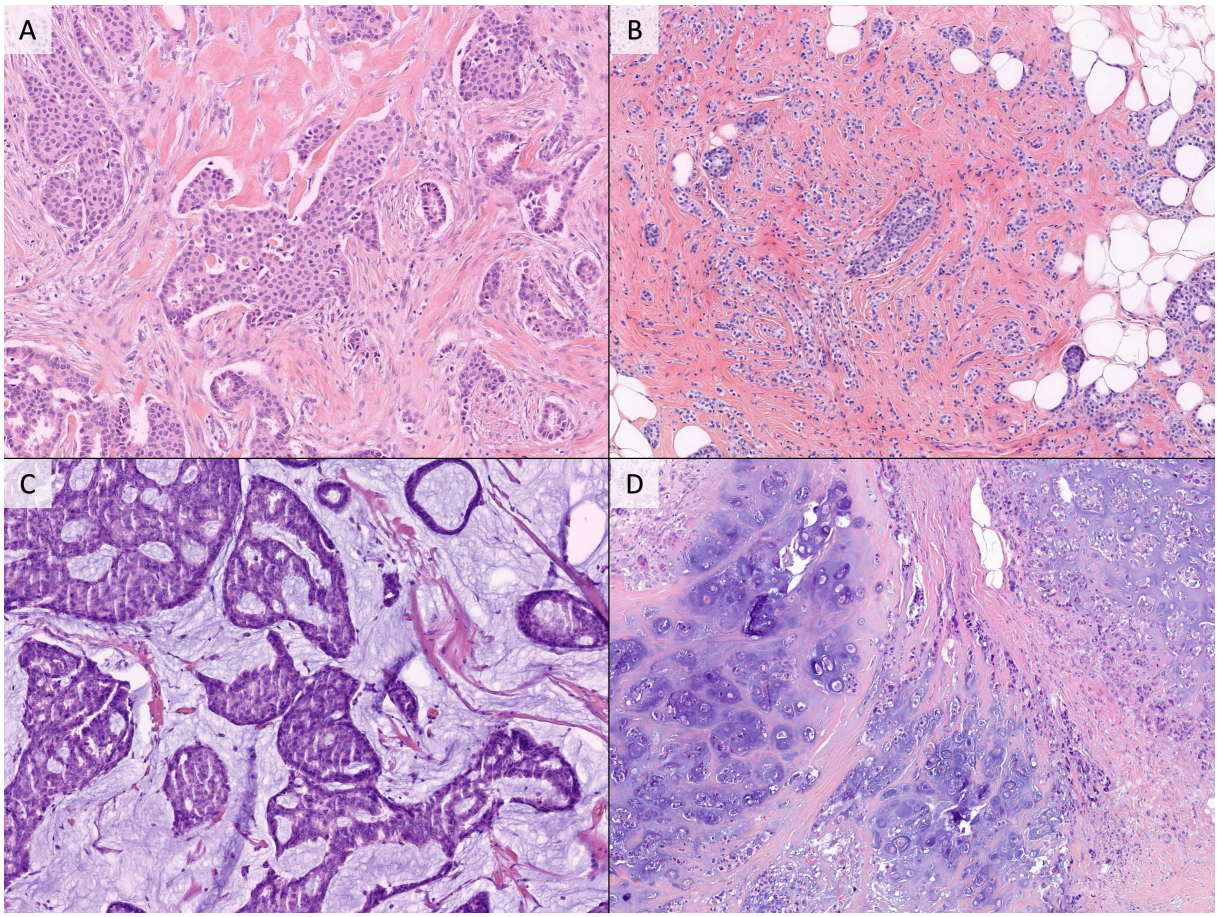


Figure 10 : Différents sous-types histologiques (grossissement x100). A. Carcinome infiltrant de type non spécifique ; B. Carcinome lobulaire infiltrant ; C. Carcinome infiltrant mucineux ; D. Carcinome infiltrant métaplasique chondroïde.

D.6.2 STADE PTNM

Il s'agit du stade tumoral évalué par le pathologiste. Le stade tumoral traduit l'extension tumorale loco-régionale et à distance. Le rôle du pathologiste consiste à mesurer la taille tumorale microscopique et à quantifier le nombre de ganglions métastatiques. Au stade précoce, l'envahissement ganglionnaire et la taille tumorale constituent les deux facteurs pronostiques les plus importants [37,38].

D.6.3 GRADE HISTOLOGIQUE

Le grade histologique a pour but de refléter l'agressivité biologique de la tumeur. Compte tenu de son importance pronostique en pratique quotidienne et de sa place fondamentale dans ce travail de thèse, la notion de grade histologique va être abordée de manière précise afin de mettre à disposition tous les éléments utiles à la discussion.

a. Histoire du grade histologique dans les cancers du sein

Afin de bien comprendre le cheminement qui a abouti à la méthode actuelle d'évaluation du grade des cancers du sein et à la notion même de grade histologique, il est intéressant d'aborder cette notion sous un angle historique. Dès 1893, David Paul von Hansemann, l'inventeur des termes « anaplasie » et « dédifférenciation », avait émis l'hypothèse que les tumeurs du sein présentant une perte de différenciation avaient une plus grande capacité de croissance [39]. Il confirmera cette hypothèse en 1902 en statuant qu'à de rares exceptions, plus le degré d'anaplasie est élevé, plus le risque de métastase était important [40]. C'était la première fois qu'un lien entre l'aspect microscopique des cellules cancéreuses mammaires et le pronostic était établi. En effet, jusqu'alors et pendant encore plusieurs décennies, seule la valeur pronostique du stade clinique (taille tumorale, envahissement ganglionnaire et présence de métastases) était prise en compte dans le choix des traitements. En 1922, MacCarthy and Sistrunk montraient une corrélation entre la survie post-mastectomie et le 1) degré de différenciation de la tumeur, 2) l'infiltration lymphocytaire et 3) l'aspect hyalin de la tumeur, pris individuellement ou en association [41]. Cette notion d'association entre infiltration lymphocytaire et pronostic fut confirmée par Alan C Perry en 1925 [42]. Alors qu'il étudiait la valeur pronostique de différents modes de croissance et d'infiltration des cellules cancéreuses mammaires, il avait constaté que la caractéristique la plus frappante était la grande proportion de patientes présentant des carcinomes médullaires (riches en lymphocytes) qui survivaient plus de 7 ans. A l'inverse, il n'avait pas retrouvé de lien évident entre l'aspect fibro-hyalin des tumeurs et un mauvais pronostic. Il est intéressant de constater, avec le recul, que la valeur pronostique des lymphocytes intra-tumoraux avait été identifiée il y a plus de 100 ans dans les cancers du sein, avant d'être longtemps négligée puis de refaire surface aujourd'hui. Toujours en 1925, Greenough a été le premier à proposer un système de grading des cancers du sein ressemblant beaucoup à celui que nous utilisons encore aujourd'hui [43]. Il proposait dès lors une classification en trois niveaux d'agressivité, basée sur le degré de différenciation tubuleuse, la taille des cellules tumorales et des noyaux, et l'hyperchromatisme et les mitoses et sur l'abondance des sécrétions. En effet, Greenough ne considérait pas que la fibro-hyalinose et l'infiltration par des petites cellules rondes (les lymphocytes) était suffisamment corrélée au niveau d'agressivité. Sur la base de sa méthode, il concluait que les patientes avec des tumeurs de grade I étaient celles qui atteignaient le plus haut pourcentage de guérison : respectivement 82% et 50% pour les stades I et II, contre 0% de guérison pour les patientes avec des tumeurs de grade III quel que soit le stade. En 1928, DH Patey et RW Scarff confirmaient l'intérêt du grade histologique de Greenough et surtout le simplifiaient pour ne prendre en compte que la proportion de formations tubuleuses (incluant la notion de secrétions), la taille des noyaux des cellules tumorales et l'hyperchromatisme (incluant les mitoses) [44]. En 1933, Haagensen

proposait à nouveau un système de grading plus complexe associant 6 critères, dont l'infiltration lymphocytaire et le caractère mucineux (*gelatinous degeneration*), finalement tombé dans l'oubli [45]. C'est la méthode initiée par Greenough, reprise et simplifiée par Patey et Scarff et finalement affinée par HJG Bloom en 1950 et HJG Bloom et WW Richardson en 1957 qui tira son épingle du jeu [46–48]. L'apport décisif de Bloom et Richardson a été de définir un score de 1 à 3 pour chaque item (formation de tubules, pléomorphisme nucléaire et mitoses) et de considérer la somme de ces scores pour différencier les tumeurs de grade I à III (Figure 11).

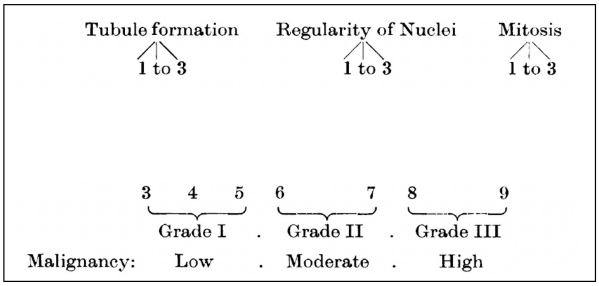


Figure 11 : Grade histologique selon Bloom et Richardson (adapté de [40])

Même si d'autres approches furent publiées dans l'intervalle, l'événement marquant qui a permis d'aboutir au grade histologique dans sa forme actuelle est la modification du score de Bloom et Richardson par CW Elston et IO Ellis en 1991 [49]. C'est, jusqu'à ce jour, l'ultime ajustement de l'évaluation du grade histologique qui a consisté en une caractérisation semi-quantitative plus précise de chaque item, avec notamment la prise en compte de l'indice de champ des microscopes (proportionnel à la surface du champ au fort grossissement) pour le score de mitoses. Cette méthode de grading, détaillée ci-dessous, peut être rencontrée sous différentes dénominations : grade de Elston et Ellis, grade de Scarff-Bloom-Richardson modifié, grade de Nottingham, grade histopronostique ou bien grade histologique (terme utilisé dans ce manuscrit).

b. Méthode d'évaluation du grade histologique

Le grade histologique peut théoriquement être évalué sur tout type de carcinome mammaire infiltrant. Il doit être évalué par un pathologiste entraîné, sur une lame en coloration standard (hématoxyline éosine) représentative de la tumeur. Les trois items (formations tubuleuses, pléomorphisme nucléaire et compte mitotique) reçoivent chacun un score entre 1 et 3.

Formations tubuleuses / glandulaires : caractérisées par une couche de cellules polarisées autour d'une lumière centrale bien visible (Figure 12).

- Score 1 : > 75% de formations tubuleuses ;
- Score 2 : 10–75% de formations tubuleuses ;

- Score 3 : < 10% de formations tubuleuses.

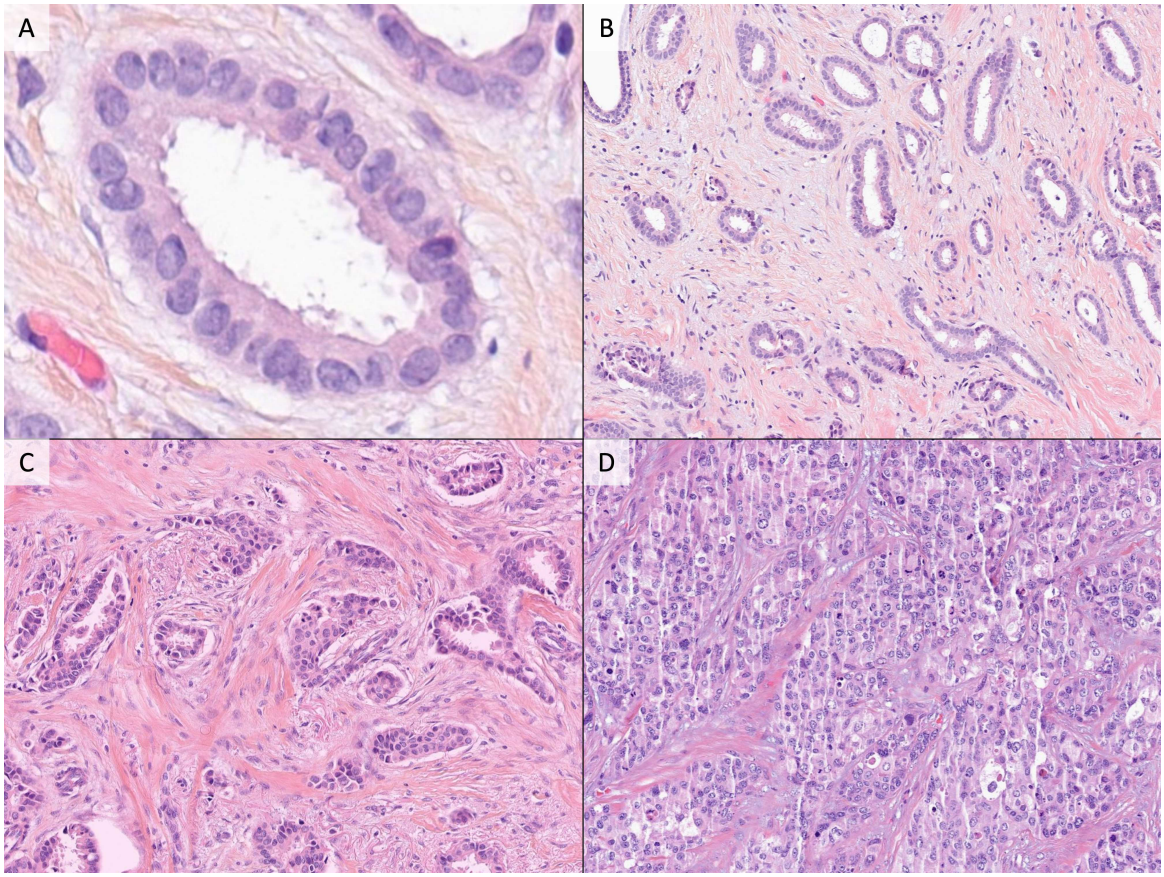


Figure 12 : Formations tubuleuses / glandulaires. A. Une formation tubuleuse est caractérisée par une couche de cellules polarisées autour d'une lumière centrale bien visible (grossissement x600). B. Score 1 : > 75% de formations tubuleuses (grossissement x 200) ; C. Score 2 : entre 10 et 75% de formations tubuleuses (grossissement x 200) ; D. Score 3 : < 10% de formations tubuleuses (grossissement x 200).

Pléomorphisme nucléaire : reflète la taille des noyaux et l'anisocaryose (Figure 13). Il doit être évalué dans la zone la moins différenciée de la tumeur. Il consiste à examiner la régularité de taille des noyaux et à la comparer à la forme des cellules épithéliales mammaires normales dans le tissu environnant.

- Score 1 : noyaux identiques en taille aux cellules épithéliales mammaires normales avec pléomorphisme minimal, chromatine fine et nucléoles discrets ;
- Score 2 : noyaux 1,5 à 2 fois plus grands que ceux des cellules épithéliales mammaires normales avec pléomorphisme modéré et nucléoles discrets ;
- Score 3 : noyaux plus de 2 fois plus grands que ceux des cellules épithéliales mammaires normales avec pléomorphisme marqué, chromatine vésiculaire et nucléoles souvent proéminents.

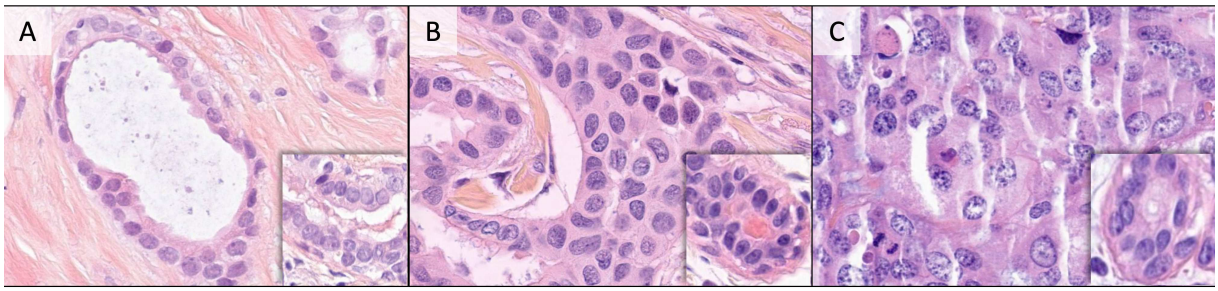


Figure 13 : Pléomorphisme nucléaire (grossissement x 400). A. Score 1 : noyaux de taille comparable à ceux des cellules épithéliales mammaires normales ; B. Score 2 : noyaux de taille légèrement augmentée avec anisocaryose modérée ; C. Score 3 : noyaux très augmentés en taille et pléomorphisme marqué. *Encarts : noyaux de glandes mammaires normales issus des mêmes lames au même grossissement pour comparaison.*

Compte mitotique : réalisé dans la zone la plus proliférante, correspondant généralement à une zone massive au front invasif de la tumeur (Figure 14). Les figures mitotiques sont quantifiées au fort grossissement (x400) sur 10 champs consécutifs, exprimé en nombre de mitoses / mm^2.

- Score 1 : < 3,65 mitoses / mm^2 ;
- Score 2 : 3,65 à 7,30 mitoses / mm^2 ;
- Score 3 : > 7,30 mitoses / mm^2.

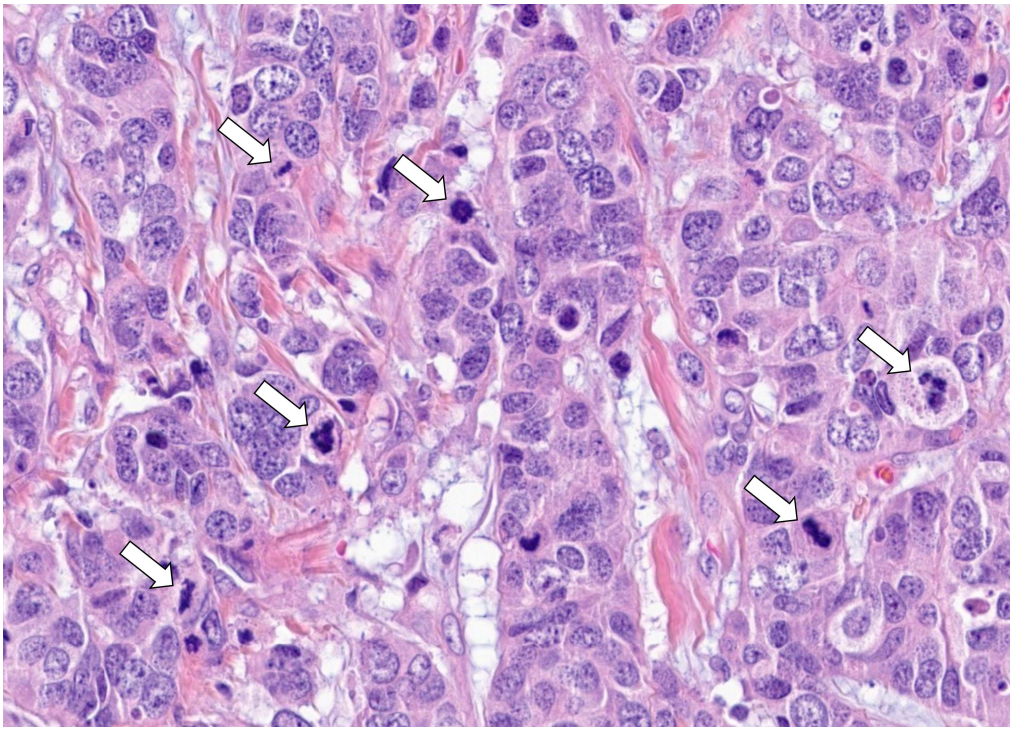


Figure 14 : Nombreuses mitoses dans un carcinome infiltrant triple négatif (grossissement x400).

Finalement, le grade histologique : se base sur la somme des trois items.

- Somme égale à 3, 4 ou 5 : grade I ;

- Somme égale à 6 ou 7 : grade II ;
- Somme égale à 8 ou 9 : grade III.

Dans la publication de Elston et Ellis, la proportion de grades I, II et III était de 19 %, 34 % et 47 % respectivement [49]. Cependant, la mise en place du dépistage organisé du cancer du sein est à l'origine d'une plus grande proportion de grades I et II dans la pratique actuelle.

c. Valeur pronostique du grade histologique

La valeur pronostique du grade histologique a souvent été remise en question. Cependant, en 2008, une étude de EA Rakha à partir des données de 2 219 patientes a permis de confirmer cette valeur pronostique, tant en termes de survie sans maladie qu'en termes de survie spécifique au cancer du sein [27]. Les courbes de survie des patientes avec des tumeurs de différents grades sont présentées dans la figure. Dans cette étude, Rakha montrait également par analyse multivariée que le grade était un facteur pronostique indépendant aussi bien chez les patientes N+ que chez les patientes N-.

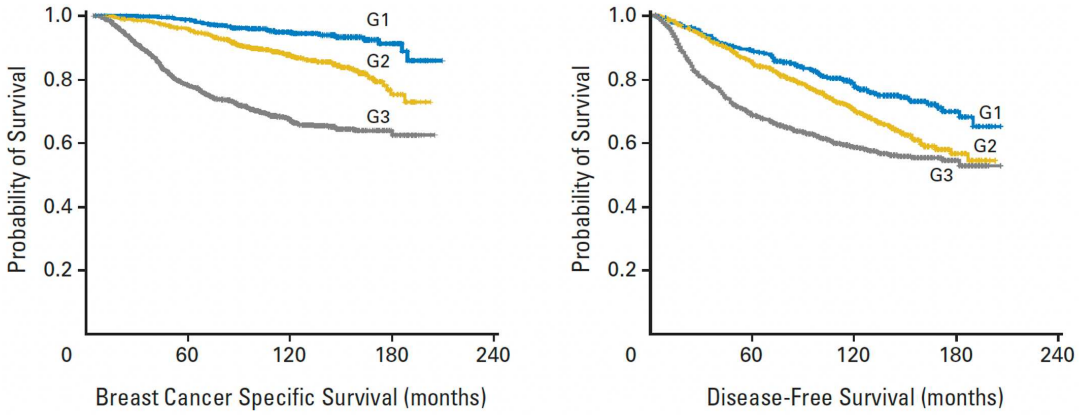


Figure 15 : Relation entre le grade histologique et, d'une part, la survie spécifique au cancer du sein (A), d'autre part, la survie sans maladie (B) [27]

D.6.4 LES EMBOLES VASCULAIRES PERI-TUMORAUX

Il s'agit de la présence de cellules carcinomateuses dans les espaces vasculaires (le plus souvent lymphatiques, mais aussi sanguins) situées à proximité de la tumeur. Ces images sont systématiquement recherchées à l'examen microscopique, car leur présence est un facteur de mauvais pronostic indépendant, tant dans la population de patientes N- (facteur de risque de rechute métastatique) que de patientes N+ (facteur de risque de rechute locale) [50,51].

D.6.5 LES RECEPTEURS HORMONAUX

L'expression par les cellules tumorales des récepteurs hormonaux (RH) (récepteurs des œstrogènes (RE) et de la progestérone (RP)) est systématiquement évaluée par immunohistochimie devant tout diagnostic de cancer infiltrant du sein. Environ 80 % des cancers du sein expriment les RE et 60 à 70% expriment les RP, 20% environ des tumeurs étant négatives pour les deux récepteurs [52,53]. Cette évaluation est surtout liée au caractère prédictif de la réponse à l'hormonothérapie de l'expression du RE. En effet, la valeur pronostique du statut des RH, même si elle est reconnue de façon consensuelle, est de moindre importance en comparaison de sa valeur prédictive. Néanmoins, il est clairement établi que la co-expression du RE et du RP est un facteur de bon pronostic, et ce d'autant plus que le niveau d'expression est élevé [53,54]. A l'opposé, une absence totale d'expression de ces deux récepteurs est corrélée à un pronostic défavorable [55,56]. Par ailleurs, une expression dissociée des récepteurs (profil phénotypique tumoral RE+ RP-) est également de moins bon pronostic, comme le montre une étude sur plus de 13 000 patientes publiée en 2005 : les tumeurs RE+/RP- sont plus proliférantes, plus souvent aneuploïdes, expriment trois fois plus l'EGFR et surexpriment plus souvent HER2 que les tumeurs RE+/RP+ [54]. La négativité du RP est associée à une plus grande hormonorésistance [57]. Les cancers RE+ sont appelés « Luminaux ». Il existe deux types de tumeurs luminales : les tumeurs luminales A, peu proliférantes et coexprimant souvent le RE et le RP, et les tumeurs luminales B, plus proliférantes, perdant souvent l'expression du RP et de plus mauvais pronostic [58,59].

Le résultat est apprécié par le pathologiste de façon semi-quantitative, par évaluation du pourcentage de cellules marquées et de l'intensité du marquage (faible, modérée, forte). Le seuil de positivité retenu en Europe est de 10% de cellules tumorales marquées (quelle que soit l'intensité du marquage). Tout carcinome infiltrant présentant au moins 10% de cellules marquées est considéré comme positif et rentre donc dans les indications de traitement par hormonothérapie adjuvante.

D.6.6 HER2

Avant les années 2000, l'évaluation du statut HER2 dans les cancers du sein avait un intérêt purement pronostique. Les tumeurs surexprimant HER2 étaient associées à un pronostic défavorable de par leurs rechutes précoces et leur évolution métastatique rapide [60–62]. A l'heure actuelle, le statut HER2 a surtout une valeur prédictive très importante de réponse aux différentes molécules anti-HER2.

En France, le statut HER2 est évalué selon les recommandations du GEFPICS [63]. Cette

évaluation fait appel à l'immunohistochimie dans tous les cas, puis à l'hybridation in situ dans les cas de statut intermédiaire par immunohistochimie. Environ 12 % des cancers du sein présentent une surexpression et/ou une amplification de HER2 [60,62]. Les taux de surexpression de HER2 sont différents en fonction du stade, de la taille, du grade, du statut ganglionnaire axillaire et du type histologique du carcinome infiltrant. Le taux de surexpression moyen (scores 3+ et 2+ amplifiés) des tumeurs T1a-T2 de moins de 3 cm est de 9 à 13% [64–66].

Une tumeur qui n'exprime ni les RH, ni HER2 est dite « triple négative ». Sauf exception, les cancers du sein triple négatif ont un comportement agressif et sont de mauvais pronostic. L'arsenal thérapeutique pour ce type de tumeur était historiquement très pauvre mais s'est considérablement étoffé ces dernières années avec l'apparition des conjugués drogue-anticorps et de l'immunothérapie.

D.6.7 INDEX DE PROLIFERATION KI67

La prolifération tumorale joue un rôle central dans l'évaluation pronostique des cancers du sein. En pratique courante, la prolifération cellulaire est évaluée par deux approches :

- Le compte mitotique sur 10 champs de surface tumorale à fort grossissement, rapporté en nombre de mitoses/mm^2, qui fait partie du grade histologique. Le compte doit être effectué au niveau des zones les plus mitotiques. La technique du compte mitotique est standardisée selon les recommandations de Elston et Ellis et la reproductibilité du compte mitotique est bonne [49].

- L'immunomarquage des cellules en cycle avec l'anticorps anti-Ki67. Ki67 est une protéine nucléaire et nucléolaire codée par le gène *MKI67* sur le chromosome 10q25. Elle est impliquée dans les phases précoces de la synthèse des ARN ribosomaux dépendante de la polymérase I et est exprimée dans les phases G1, S, G2 et M du cycle cellulaire *[45]*. Son expression est détectée par immunohistochimie. Le groupe de travail international sur le Ki67 dans les cancers du sein estime qu'un Ki67 ≤ 5 % est gage de faible prolifération et de bon pronostic alors qu'un Ki67 ≥ 30 % est de mauvais pronostic [67]. Entre ces deux bornes, le Ki67 ne devrait pas guider la décision de chimiothérapie. Cependant, dans certaines études récentes, un seuil à 20% était encore considéré comme un facteur associé à un haut risque de rechute [68]. Le Ki67 reste un facteur pronostique indépendant de niveau de preuve Ib selon les résultats de vastes méta-analyses incluant plus de 10 000 patientes [69–71].

E. STRATEGIE DE TRAITEMENT EN PHASE PRECOCE

En pratique courante, la stratégie de traitement des cancers du sein infiltrants repose sur le stade TNM clinique (cTNM), le statut ménopausique, le pTNM, le grade histologique, l'immunophénotype (récepteurs hormonaux et HER2) et l'index de prolifération Ki-67.

Dans la très grande majorité des cas, les tumeurs de grade I expriment le RE et les patientes sont donc éligibles à une hormonothérapie seule, sans chimiothérapie.

À l'autre extrémité du spectre, les tumeurs de grade III peuvent présenter des immunophénotypes variés, mais le traitement comprendra presque toujours de la chimiothérapie.

Entre les deux se situent les patientes avec des tumeurs de grade II, pour qui la décision de chimiothérapie reste beaucoup plus difficile à prendre, notamment pour les patientes dont la tumeur exprime le récepteur des œstrogènes. En effet, sauf exception, les tumeurs triple négatives seront traitées par chimiothérapie ainsi que les tumeurs HER2+ qui recevront du trastuzumab, toujours donné en association avec de la chimiothérapie.

Pour les patientes avec une tumeur de grade II RE+ HER2- au stade localisé, ce qui représente environ 40 % des patientes atteintes de cancer du sein, la décision de chimiothérapie est donc discutée en RCP au cas par cas. Sur la base d'un faisceau d'arguments associant l'envahissement ganglionnaire, le statut ménopausique et l'évaluation de la prolifération tumorale (Ki67) on tente de faire pencher la balance bénéfice-risque d'un côté ou de l'autre. Cependant, dans certains cas, le recours à une signature d'expression génique peut être décidé afin d'orienter la décision et notamment d'aider à la décision d'abstention de chimiothérapie [72,73]. Il n'est pas nécessaire de détailler ici le fonctionnement de ces signatures d'expression génique, mais simplement rappeler qu'elles sont coûteuses et leur remboursement par la sécurité sociale n'est pas encore assuré. La figure décrit les principales caractéristiques des signatures d'expression génique développées depuis les années 2010. C'est cette fréquente situation d'indécision clinique, aboutissant à la prescription de tests coûteux, qui était à l'origine du projet de thèse initial.

Nom générique	70-*gene signature*	21-*gene signature*	PAM50	97-*gene genomic grade* simplifié 8 gènes	5-*gene molecular grade index* + HOXB13:IL17BR 2-*gene ratio*	11-*gene assay*
Nom commercial	**MammaPrint™**	**Oncotype DX®**	**Prosigna**	**MapQuant DX (97 gènes)**	***Breast Cancer Index***	**Endopredict®**
Fournisseur	Agendia BV (Amsterdam, Pays-Bas)	Genomic Health (Redwood City, CA, USA)	Nanostring Technologies, Inc., Seattle, WA, USA	Ipsogen SA et Halio Dx (Marseille, France et Stamford, CT, USA)	bioTheranostics, Inc. (San Diego, CA, USA)	Sividon diagnostics GmbH (Koln, Germany)
Indication principale	CSI de stade I/II, ≤ 61 ans, diamètre < 5 cm, N0, RE- ou RE+	CSI RE+, N0 Prédiction de réponse à la CT adjuvante	CSI N0 ou N+ Classification moléculaire	CSI de grade II histologique, RE+	CSI RE+, N0	CSI RE+, HER2-
Méthode	*Microarray* (Agilent Technologies, Inc., Santa Clara, CA, USA)	qRT-PCR	qRT-PCR nCounter Analysis System (Nanostring Technologies, Inc., Seattle, WA, USA)	Microarray (Affymetrix, Santa Clara, CA, USA) (97 gènes)/qRT-PCR (simplifié)	RT-PCR	qRT-PCR
Résultats du test	Haut risque/bas risque	*Recurrence score* 0 à 100 Haut risque/risque intermédiaire/bas risque	Sous-type moléculaire *Risk of recurrence score* 0 à 100 Haut risque/ risque intermédiaire/ bas risque	Grade génomique : haut grade (GG-3)/ bas grade (GG-1)/ équivoque (GG-2, catégorie rajoutée secondairement)	0 à 10 Haut risque/risque intermédiaire/bas risque	0 à 15 Haut risque/bas risque
Gènes associés à la biologie du cancer du sein	Non détaillés	*ER, PR, BCL2, SCUBE2, Ki67, STK15, Survivin (BIRC5), CCNB1, MYBL2, HER2, GRB7, MMP11, CTSL2, GSTM1, CD68, BAG1*	50 gènes (non détaillés)	97 gènes (non détaillés) Simplifié : *MYBL2, KPNA2, CDC2, CDC20*	*HOXB13, IL17BR, BUB1, CENPA, NEK2, RACGAP1, RRM2*	*DHCR7, AZGP1, MGP, STC2, BIRC5, UBE2C, RBBP8, IL6ST*
Gènes de référence	Non détaillés	*GAPDH, ACTB, RPLPO, GUS, TFRC*	8 gènes (non détaillés)	97 gènes : non détaillés Simplifié : *GUS, TBP, RPLPO, TFRC*		*CALM2, OAZ1, RPL37A*
Échantillon tissulaire	Congélation ou fixé en formol et inclus en paraffine	Fixé en formol et inclus en paraffine	Fixé en formol et inclus en paraffine	Congélation (97 gènes)/fixé en formol et inclus en paraffine (simplifié)	Fixé en formol et inclus en paraffine	Fixé en formol et inclus en paraffine

Figure 16 : Récapitulatif des classifications moléculaires pronostiques pour les cancers du sein de sous-type luminal (adapté de [73]).

F. PROJET DE RECHERCHE INITIAL ET PREMIERS RESULTATS

F.1 HYPOTHESE DE RECHERCHE

Une analyse d'image performante pourrait identifier, au sein d'images microscopiques de cancer du sein de grade II RE+ HER2-, des informations pronostiques pertinentes et complémentaires des facteurs pronostiques morphologiques déjà connus. Ceci permettrait de guider la décision de chimiothérapie adjuvante tout en limitant éventuellement le recours aux signatures d'expression génique.

F.2 DESIGN DE L'ETUDE

Le design de l'étude s'apparentait à celui utilisé pour le développement de la signature moléculaire Genomic Grade Index développée par l'équipe de Christos Sotiriou [74,75]. En effet, le développement de cette signature moléculaire se basait sur la création d'un classifieur de cancers du sein grade I versus grade III à l'échelle transcriptomique. Une fois entraîné et

atteignant de bonnes performances dans la distinction des tumeurs de grade I et de grade III, le transcriptome de tumeurs de grade II était soumis au modèle afin de les classer en *grade I-like* ou en *grade III-like*. De manière intéressante, cette signature moléculaire permettait réellement de séparer les tumeurs de grade II en termes de pronostic sur la base de leur transcriptome.

Ce design pouvait facilement être adapté à un contexte d'images microscopiques selon le plan suivant :

1) entraînement d'un classifieur grade I versus grade III ;

2) validation des performances sur cette tâche de classification ;

3) utilisation du classifieur pour séparer les tumeurs de grade II en *grade I-like* ou en *grade III-like* ;

4) comparaison des pronostics des patientes dans ces deux groupes.

Un intérêt majeur de ce design est qu'il conserve toutes les données des patientes avec tumeur de grade II, et donc tous les événements cliniques (rechute, évolution métastatique), pour la phase de validation du modèle. L'analyse de la valeur pronostique du modèle profite donc du maximum de puissance statistique.

F.3 COHORTES

Dans le cadre de ce travail, nous avons eu accès à trois cohortes nationales de patientes atteintes d'un cancer du sein pour lesquelles une lame représentative de leur cancer était disponible. Ces cohortes provenaient de différentes études du programme PACS (Programme adjuvant dans le cancer du sein) de la fédération des Centres de Lutte Contre le Cancer - UNICANCER. Il s'agissait de trois études dont les résultats étaient négatifs, ce qui signifie qu'il n'existait pas de différence significative en termes de pronostic dans les bras de traitement de chacune de ces études. Elles sont succinctement détaillées ci-dessous, avec un focus particulier sur les critères d'inclusion et le nombre de lames disponibles.

Étude PACS04 [76,77] : l'objectif de cette étude était de comparer deux schémas de chimiothérapie : 6 cycles de FEC 100 (5Fluorouracile-Epirubicine-Cyclophosphamide) versus 6 cycles de Epirubicine-Docetaxel (ED) + trastuzumab séquentiel pour les HER2+ en adjuvant pour les patientes atteintes d'un cancer du sein HER2 positif avec envahissement ganglionnaire. Au cours de cette étude, 3009 patientes dont 528 avec une tumeur HER2+ ont été incluses et 1892 lames étaient disponibles. Les blocs FFPE de cette étude avaient été centralisés et les lames colorées à l'Institut Gustave Roussy de Villejuif

Étude PACS05 [78] : l'objectif de cette étude était de comparer deux schémas de chimiothérapie : 4 cycles de FEC100 vs 6 cycles FEC100 pour les patientes atteintes d'un cancer du sein sans envahissement ganglionnaire, avec un élément en faveur d'un « haut risque de récidive » parmi une taille tumorale > 2 cm, une négativité des RH, un grade histologique II ou III ou un âge ≤ 35 ans. Au cours de cette étude, 1515 patientes de tout immunophénotype ont été incluses et 719 lames étaient disponibles. Les blocs FFPE de cette étude avaient été centralisés et les lames colorées au Centre Jean Perrin de Clermont-Ferrand

Étude PACS08 [79] : l'objectif de cette étude était de comparer deux schémas de chimiothérapie : associant FEC100 et soit docetaxel, soit ixabepilone pour les patientes atteintes d'un cancer du sein localisé de mauvais pronostic. Étaient incluses des patientes avec tumeur triple négative avec ou sans envahissement ganglionnaire ou luminale B avec envahissement ganglionnaire. Au cours de cette étude, 762 patientes ont été incluses et 742 lames étaient disponibles. Les blocs FFPE de cette étude avaient été centralisés et les lames colorées à l'Institut Claudius Regaud de Toulouse.

En conclusion, le pronostic des patientes dans les différentes cohortes n'était pas identique. Le classement des cohortes en termes de facteurs pronostiques associés aux critères d'inclusion (statut ganglionnaire et immunophénotype) était le suivant, du moins péjoratif au plus péjoratif :

- PACS05 : patientes pN0, tout immunophénotype ;
- PACS04 : patientes pN+, tout immunophénotype ;
- PACS08 : patientes avec tumeur triple négative ou luminale B N+.

La politique d'accès aux données d'UNICANCER est stricte. Les bases de données associées aux lames ne sont jamais transférées au porteur de la recherche mais à un biostatisticien. Dans le cadre d'approche de machine learning, le biostatisticien gère l'accès aux données et se charge de conserver les données de test (de manière équilibrée entre les cohortes et sur un certain nombre de critères discutés avec le porteur de projet). Dans le cadre de notre étude, notre biostatisticien nous a transmis les identifiants de lames de grades I et III des différentes cohortes en gardant assez de lames pour la phase de test finale.

Le nombre de lames virtuelles disponibles, par cohorte et par grade, pour la phase de développement du modèle, était déséquilibré au même titre que dans les cohortes initiales. La répartition des lames utilisées pour la phase de développement est indiquée dans la Figure 17.

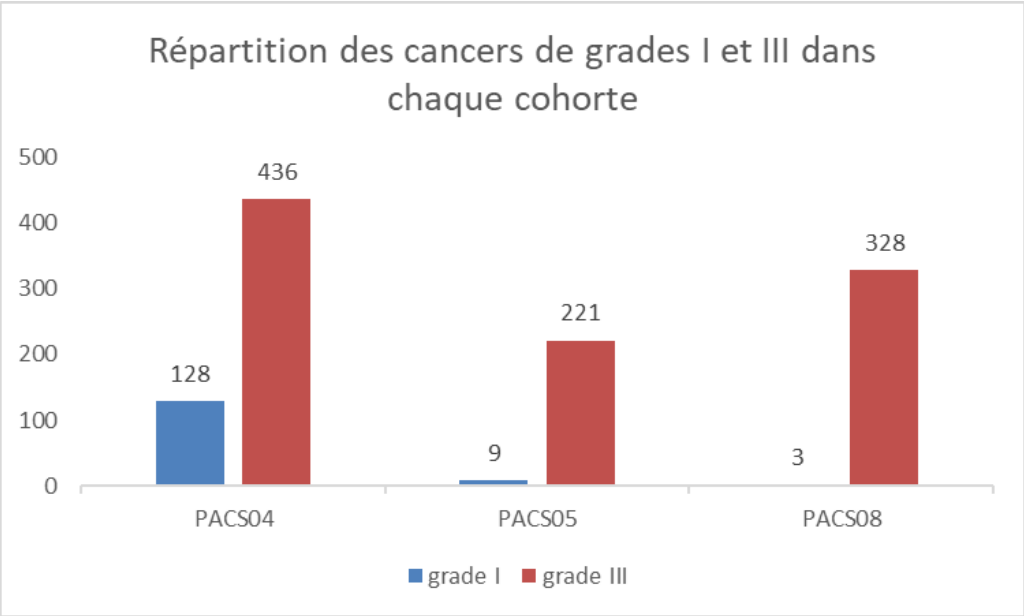


Figure 17 : Nombre de lames virtuelles par cohorte et par grade.

Toutes les lames ont été numérisées à l'IUCT-Oncopole dans le cadre de ce travail de thèse.

F.4 PREMIERS RESULTATS ET IDENTIFICATION D'UN BIAIS MAJEUR DANS LE DATASET

Une séparation du dataset en entraînement - validation - test a été réalisée à l'échelle des lames en favorisant un équilibre en termes de cohorte d'origine et de grade. Un réseau de neurones convolutif (CNN) a alors été entraîné selon la méthode décrite dans le chapitre Material and Methods de l'article. Compte tenu du déséquilibre en termes de nombre de cas dans les différentes cohortes et en termes de répartition des grades I et III, un sous-échantillonnage des groupes majoritaires a été réalisé afin d'obtenir un dataset équilibré en termes de cohortes et de grades.

Les premiers résultats obtenus étaient extrêmement prometteurs, avec une performance globale à l'échelle des patches de 86% (*accuracy*) et une aire sous la courbe ROC (AUROC) de 0,92. La matrice de confusion est présentée dans la Figure 18, ainsi que l'histogramme représentant la répartition des scores de prédiction obtenus.

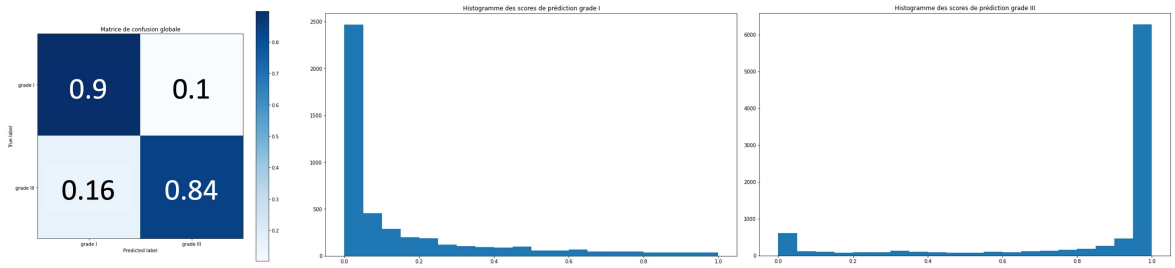


Figure 18 : Matrice de confusion et histogramme des scores de prédiction du modèle entraîné.

Malgré ces résultats encourageants, une visualisation plus précise par cohorte d'origine a permis de mettre en évidence des résultats beaucoup plus contrastés. En effet, les bonnes performances globales du modèle cachaient une tendance au sous-grading dans la cohorte PACS05 et au sur-grading dans la cohorte PACS08. Soixante pour cent des patchs de tumeurs de grade III étaient prédits en grade I dans la PACS05 et 30 % des patchs de tumeurs de grade I étaient prédits en grade III dans la PACS08. Les performances restaient bonnes sur les patchs issus de la cohorte PACS04 (figure).

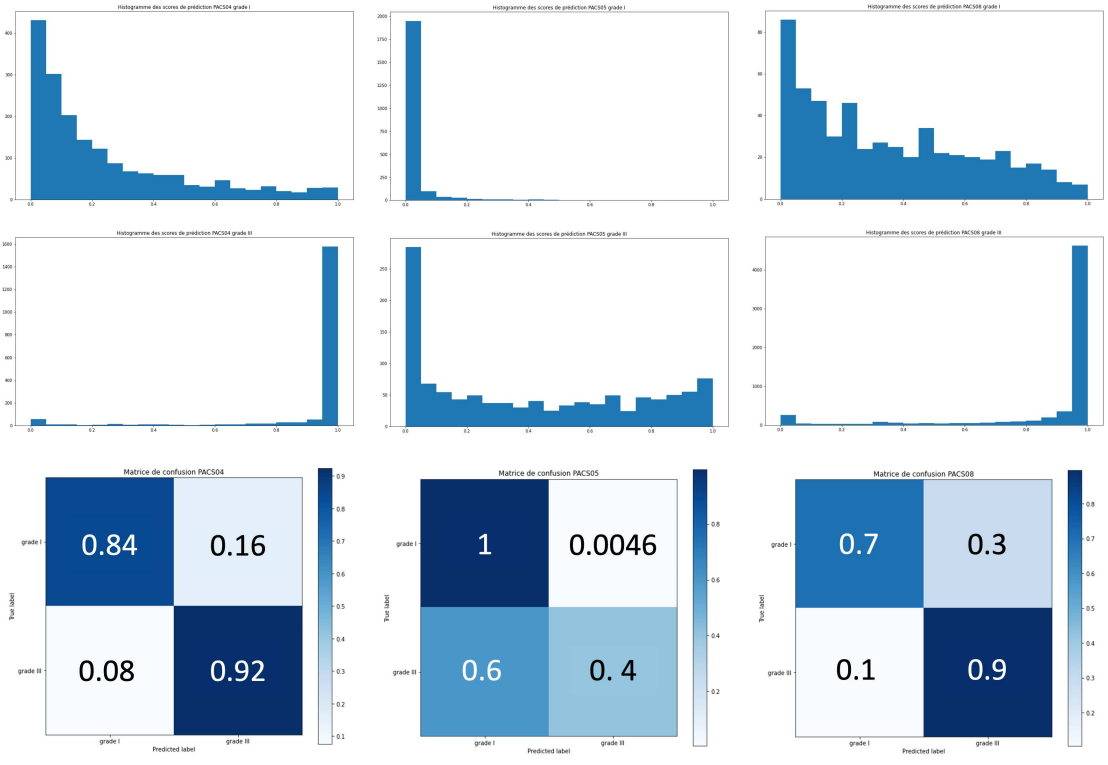


Figure 19 : Matrice de confusion et histogramme des scores de prédiction du modèle entraîné pour chaque cohorte.

Compte tenu des différences en termes de pronostic des cohortes PACS et du fait que les lames de chaque cohorte avaient été colorées dans un centre différent, nous avons émis l'hypothèse

que le réseau de neurones avait appris à séparer les cohortes notamment sur la base de la coloration. Nous avons pu étayer notre hypothèse en réalisant une représentation graphique en 2D de la dispersion de la teinte des patches des différentes cohortes (figure). Afin de visualiser l'effet de cette dispersion sur l'apprentissage du modèle, nous avons également représenté la séparation des cohortes PACS par le réseau de neurones, sous forme d'une projection orthogonale de la représentation des patchs dans l'espace latent du modèle ResNet entraîné (figure). Nous avons pu conclure que, dans notre dataset, la coloration HE était statistiquement liée aux cohortes et indirectement au grade histologique du fait des critères d'inclusion des différentes études.

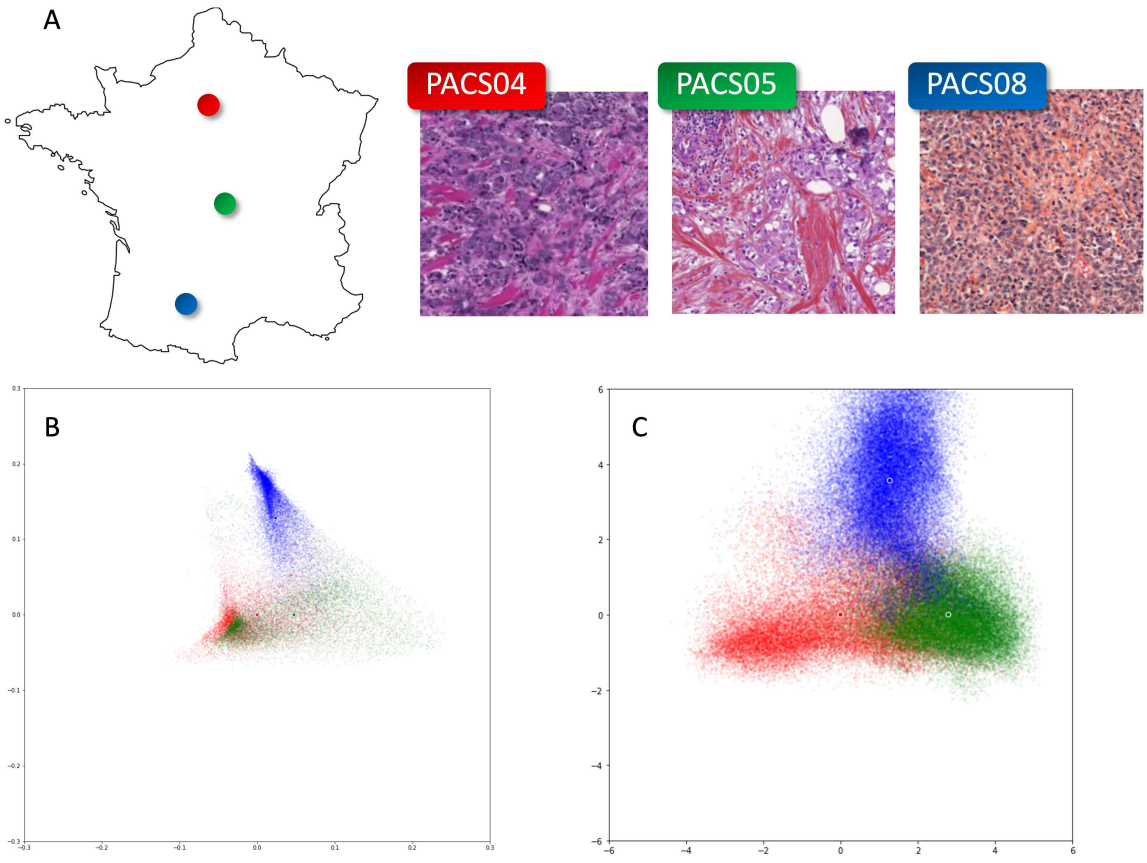


Figure 20 : Mise en évidence d'un biais lié à la coloration des lames dans les différentes cohortes. A. Images représentatives des colorations dans les différentes cohortes en fonction du site ; B. Projection orthogonale en 2D de la teinte (canal H) moyenne de chaque patch du dataset ; C. projection orthogonale en 2D de la représentation des patchs dans l'espace latent du modèle ResNet entraîné. Points rouges : PACS04, points verts : PACS05, points bleus : PACS08.

A la suite de la découverte de ce biais majeur dans nos données d'apprentissage, nous avons tenté d'utiliser des méthodes d'augmentation de données usuelles sans véritable succès sur l'atténuation du biais. Nous avons donc choisi de nous concentrer sur cet aspect pour la suite du

travail, avec pour objectif de tirer parti du dataset à notre disposition qui s'avérait être un modèle caricatural de dataset biaisé. Nous nous sommes concentrés sur le développement d'outils de normalisation et d'augmentation de données adaptés au domaine de la pathologie, en comparant nos méthodes avec les méthodes existantes dans la littérature. Nous avons profité du dataset à notre disposition pour non seulement évaluer l'efficacité des méthodes de normalisation et d'augmentation sur les couleurs, mais aussi pour monitorer leur impact en termes de généralisation d'un modèle de classification (grade I *vs* grade III).

G. INTRODUCTION SUR LA NOTION DE BIAIS

G.1 DEFINITIONS

Il existe plusieurs définitions de la notion de biais suivant que l'on s'intéresse au domaine précis du machine learning dans ses aspects mathématiques et statistiques ou au domaine plus général du biais en sociologie.

G.1.1 DEFINITION MATHEMATIQUE DANS LE CONTEXTE DU MACHINE LEARNING

La notion de biais mathématique en machine learning est indissociable de la notion de variance. Un modèle peut avoir un biais faible ou élevé et une variance faible ou élevée. Lorsque l'on développe un modèle de machine learning, on doit faire un compromis entre le biais et la variance : un modèle trop simple ne capte pas la grande tendance associée à la distribution des données d'apprentissage, ses performances en tests seront moyennes ou bonnes mais ne deviendront jamais très bonnes : il s'agit d'un *biais élevé*. Un modèle trop complexe acquiert trop de détails issus des données d'apprentissage (sur lesquels il est excellent, voire trop bon) et par conséquent, peut se comporter alternativement très bien ou très mal lors de la phase de déploiement : il s'agit d'une *variance élevée*, associée à la notion de surapprentissage. En réalité, compte tenu de la complexité potentiellement gigantesque (et toujours croissante) des réseaux de neurones artificiels, ce que nous avons appelé « biais » jusqu'ici correspondait plutôt à un phénomène de surapprentissage lié à une variance élevée. Les principaux axes stratégiques pour trouver un bon compromis biais-variance consiste à tester des modèles de complexité croissante et à utiliser des approches de pénalisation (bagging, boosting, régularisation). Un bon reflet de la complexité d'un modèle est son nombre de paramètres. Une régression linéaire simple ajustera 2 paramètres, là où les réseaux de neurones artificiels les plus récents peuvent dépasser le milliard de paramètres.

G.1.2 DEFINITION SOCIOLOGIQUE

Dans ce travail de thèse, c'est plutôt dans son acception sociologique que l'on aborde cette notion de biais : une tendance pour un algorithme à introduire une erreur systématique dans le processus d'apprentissage qui nuit à la validité des prédictions qui en résultent[11]. La notion d'erreur systématique est fondamentale. Par exemple, dans la description du biais initial de notre modèle, on voit que l'erreur n'était pas aléatoire, elle semblait orientée dans un sens pour la PACS05 et dans l'autre pour la PACS08, faisant écho au fait qu'il existait un lien direct entre la coloration et le pronostic dans notre dataset. C'est ce caractère systématique qui rend le modèle imprévisible et les prédictions hasardeuses lors de l'étape de déploiement du modèle.

G.2 SOURCES DE BIAIS EN PATHOLOGIE

Il existe des sources de biais évidentes en pathologie [80]. La coloration des lames est probablement la plus évidente du fait de la variabilité technique évoquée précédemment. Au sein d'un même laboratoire, la coloration peut aussi poser problème notamment lors de l'utilisation de lames d'ancienneté différente. Outre le fait que la technique de coloration ait pu changer dans le temps dans un laboratoire, c'est surtout la modification progressive de la coloration HE au cours du temps qui rend « détectable » l'ancienneté d'une lame. L'acquisition des images est également une source potentielle de biais. En effet, chaque capteur a sa propre sensibilité et sa propre résolution. A objectif identique et pour une lame identique, le rendu final des lames virtuelles ne sera pas exactement le même en fonction du scanner utilisé. Enfin, un biais beaucoup plus difficile à identifier concerne la disparité d'accès à la numérisation des images en routine. Tous les laboratoires de pathologie ne sont pas équipés de scanners de lames. Il est probable que, de manière statistique, les grands centres ou les grands groupes, dont le recrutement n'est probablement pas identique aux plus petits laboratoires, soit plus équipés et soit donc plus en mesure de construire des datasets de WSI, instillant un biais sous-jacent dont l'impact est difficile à évaluer. Par exemple, il a été prouvé que les lames virtuelles (WSI) du Cancer Genome Atlas (TCGA) étaient biaisées. En effet, dans cette étude menée par Taher Dehkharghanian et collaborateurs, des modèles de deep learning pouvaient prédire le site d'origine des WSI avec une *accuracy* située entre 70 % et 86 % en identifiant un pattern morphologique spécifique de chaque site [81]. Ils ont également montré qu'un tel pattern, bien que médicalement non pertinent, pouvait avoir un rôle non négligeable dans les prédictions d'un modèle de classification de sous-type tumoral. Ces éléments vont dans le sens d'un précepte bien connu : « *AI can learn without being taught* ». Et surtout, dépourvue de bon sens, l'IA n'a aucun

[11] définition adaptée du document *Doing Global Science: A Guide to Responsible Conduct in the Global Research Enterprise*, InterAcademy Partnership, Princeton University, 2016. page 6 : "(…) a bias is a tendency or inclination on the part of a researcher or research group that intricues systematic error into the research process and damages the validity of the resulting work."

moyen de faire le tri entre des concepts médicalement pertinents ou non. Dans notre exemple, il serait exagéré de penser que l'algorithme a appris les critères d'inclusion des études à partir des images, cependant il semble avoir identifié une corrélation entre un élément médicalement non pertinent (la coloration des lames) et la classe à prédire (le grade).

Dans leur article de revue, K Nakagawa et collaborateurs identifient en priorité deux méthodes d'atténuation du biais lié à la préparation des lames : la normalisation et l'augmentation de données. C'est également sur ces deux points que nous avons porté notre attention dans le cadre de ce travail.

G.3 SOURCES DE BIAIS DANS LES ETUDES CLINIQUES

Il existe un grand nombre de biais potentiels associés aux essais cliniques. On comprend aisément l'impact que peuvent avoir les critères d'inclusion et d'exclusion, les méthodes de mesure ainsi que la part de subjectivité qui associe chaque participation (ou non-participation) à un essai.

- Biais de sélection : différences systématiques entre les caractéristiques de base des groupes comparés.
- Biais de performance : différences systématiques dans la façon dont les groupes sont traités dans l'expérience, ou exposition à d'autres facteurs que l'intervention d'intérêt.
- Biais de détection : différences systématiques entre les groupes quant à la façon dont les résultats sont déterminés.
- Biais d'attrition : différences systématiques entre les groupes dans les retraits d'une étude, ce qui peut mener à des résultats sur des données incomplètes.
- Biais de déclaration : différences systématiques entre les constatations déclarées et les constatations non déclarées.
- Biais de l'expérimentateur : biais subjectif en faveur d'un résultat attendu par l'expérimentateur.
- Biais de confirmation : la tendance à rechercher, interpréter ou favoriser l'information d'une manière qui confirme ses idées préconçues ou hypothèse.[12]

[12] Source: The Academy of Medical Science, Reproducibility and reliability of biomedical research : improving research practice. Symposium report, October 2015. https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Lorsque l'on accède aux données d'un ou plusieurs essais cliniques pour le développement de modèles de machine learning, la connaissance précise du protocole de recherche, de la méthode et des résultats permet d'apprécier les limites potentielles des algorithmes qui en découlent. Le biais de notre modèle, si nous devions faire le parallèle avec un essai clinique, correspondrait à un biais de sélection. En effet, la variation de coloration des lames n'est pas, en elle-même, problématique. C'est le lien systématique entre cette coloration et les critères d'inclusion des études qui crée le biais.

G.4 SOURCES DE BIAIS EN MACHINE LEARNING

Les sources de biais en machine learning sont nombreuses. Elles se divisent en deux grandes catégories : les biais provenant des données et les biais provenant de facteurs spécifiques aux modèles [82]. Dans le cadre d'une revue systématique de la littérature, Van Giffen et ses collègues ont identifié huit familles de biais en machine learning [83]. Si certains de ces biais sont directement applicables au domaine de la pathologie digitale, d'autres semblent plus théoriques.

- Biais social : les données disponibles reflètent un biais social préexistant dans la population cible. Ce type de biais a été démontré de multiples fois dans le domaine médical, y compris en pathologie [84].
- Biais de mesure : les features et les annotations (labels) choisis reflètent imparfaitement les véritables variables d'intérêt.
- Biais de représentation : les données d'entrée ne sont pas représentatives de la population concernée.
- Biais d'annotation : les annotations (labels) s'écartent systématiquement des catégories réelles sous-jacentes. Ceci inclut la subjectivité du pathologiste au cours d'une annotation manuelle d'images microscopiques. Dans notre exemple, malgré tous les efforts réalisés pour harmoniser l'évaluation du grade histologique, il reste une part de subjectivité et la concordance entre pathologistes est loin d'être parfaite [39]. Dans le contexte des essais cliniques, une relecture centralisée par un pathologiste expert est souvent réalisée afin d'assurer une concordance interne à l'étude, diminuant de ce fait les biais de sélection et/ou de détection, mais en les remplaçant par un biais d'annotation lié à la subjectivité que peut apporter un expert unique.
- Biais algorithmique : des considérations techniques inappropriées lors de la modélisation conduisent à une déviation systémique du résultat.
- Biais d'évaluation : utilisation d'une population de test non représentative ou de métriques de performance inappropriées.

- Biais de déploiement : Le modèle est utilisé et interprété dans un contexte différent de celui pour lequel il a été conçu.
- Biais de rétroaction (*feedback*) : Le résultat du modèle influence les données d'apprentissage futures, de telle sorte qu'un léger biais peut être renforcé par une boucle de rétroaction.

G.5 LA FUITE DE DONNEES

La fuite de données correspond à un biais d'évaluation, mais mérite un chapitre à part du fait du risque majeur qu'elle représente en pathologie digitale et de manière générale en médecine. Une fuite de données se produit lorsque des informations extérieures à l'ensemble de données d'entraînement influencent par inadvertance le modèle, ce qui conduit à des estimations de performance trop optimistes. De manière plus concrète, une fuite de données intervient dès lors que les données d'un même patient se retrouvent dispatchées à la fois dans le dataset d'apprentissage et dans le dataset de validation ou de test (Figure 21).

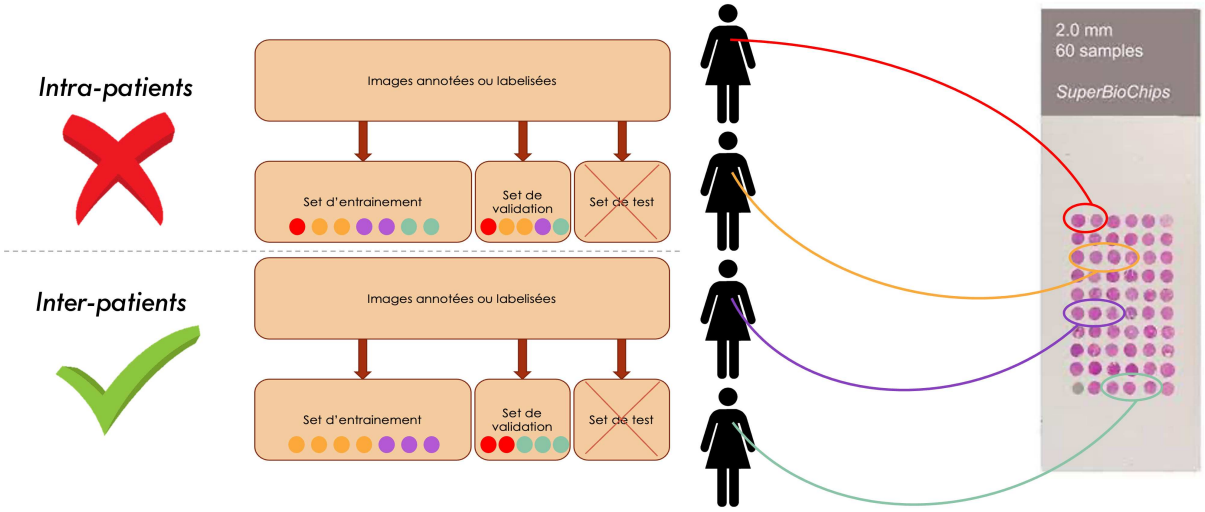


Figure 21. : Exemple de fuite de données dans un projet de machine learning. Prédiction de rechute à 3 ans de cancers du sein à partir d'images de Tissue Micro Array en coloration standard (HE). Le cancer du sein d'une même patiente peut être représenté par plusieurs carottes. Deux approches de *data splitting* étaient comparées : une approche inter-patients avec *data splitting* rigoureux et une approche intra-patients avec fuite de données. Les performances du modèle étaient de 62,4 % d'*accuracy* avec l'approche rigoureuse contre 70,3 % *d'accuracy* avec la fuite de données [85]. Les modèles n'étaient pas évalués sur un dataset de test.

Pour prévenir ce type de phénomène, il est important de faire la séparation des datasets de manière extrêmement rigoureuse et à l'échelle du patient (Figure 22). En effet, des patchs issus d'une même lame virtuelle, même s'ils intéressent une zone différente de la lame, ne doivent pas

être répartis dans les différents datasets. De la même manière, si plusieurs lames concernent le même patient, la même règle s'applique. Il s'agit d'une erreur fréquente dans les publications de machine learning, pour lesquelles la procédure de *data splitting* est souvent opaque, voire manifestement inadéquate [85].

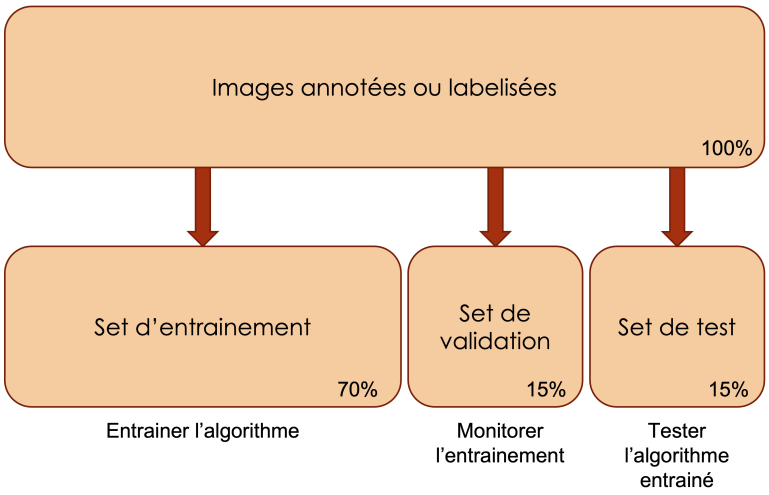


Figure 22 : Schéma général du *data splitting* en machine learning. Les proportions de chaque dataset sont indicatives et peuvent varier d'un projet à l'autre.

II. PUBLICATION

Computers in Biology and Medicine 171 (2024) 108130

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Bias reduction using combined stain normalization and augmentation for AI-based classification of histological images

Camille Franchet [a,b,*,1], Robin Schwob [c,1], Guillaume Bataillon [a], Charlotte Syrykh [a], Sarah Péricart [a], François-Xavier Frenois [a], Frédérique Penault-Llorca [d], Magali Lacroix-Triki [e], Laurent Arnould [f], Jérôme Lemonnier [g], Jean-Marc Alliot [h,i], Thomas Filleron [j], Pierre Brousset [a,b,k]

[a] *Institut Universitaire du Cancer-Oncopole, Pathology Department, F-31059 Toulouse, France*
[b] *Université Toulouse III-Paul Sabatier, F-31000 Toulouse, France*
[c] *Thales Services Numériques, Augmented data department, F-31670 Labège, France*
[d] *Centre Jean Perrin, Université Clermont Auvergne, INSERM, U1240 Imagerie Moléculaire et Stratégies Théranostiques, F-63000 Clermont Ferrand, France*
[e] *Département de Pathologie, Gustave-Roussy Cancer Campus, F-94805 Villejuif, France*
[f] *Department of Pathology and Tumor Biology, Centre Georges François Leclerc, Dijon, France*
[g] *R&D Unicancer, Paris, France*
[h] *Institut de Recherche en Informatique de Toulouse, Toulouse University, Toulouse, France*
[i] *CHU Toulouse, Toulouse, France*
[j] *Institut Universitaire du Cancer-Oncopole, Institut Claudius Regaud, Biostatistics & Health Data Science Unit, F-31059 Toulouse, France*
[k] *Laboratoire d'Excellence Toulouse Cancer-TOUCAN, UMR1037 CRCT, F-31037 Toulouse, France*

ARTICLE INFO

Keywords:
Histopathology
Bias mitigation
Color-induced bias
Data augmentation
Normalization
Deep learning

ABSTRACT

Artificial intelligence (AI)-assisted diagnosis is an ongoing revolution in pathology. However, a frequent drawback of AI models is their propension to make decisions based rather on bias in training dataset than on concrete biological features, thus weakening pathologists' trust in these tools. Technically, it is well known that microscopic images are altered by tissue processing and staining procedures, being one of the main sources of bias in machine learning for digital pathology. So as to deal with it, many teams have written about color normalization and augmentation methods. However, only a few of them have monitored their effects on bias reduction and model generalizability. In our study, two methods for stain augmentation (AugmentHE) and fast normalization (HEnorm) have been created and their effect on bias reduction has been monitored. Actually, they have also been compared to previously described strategies. To that end, a multicenter dataset created for breast cancer histological grading has been used. Thanks to it, classification models have been trained in a single center before assessing its performance in other centers images. This setting led to extensively monitor bias reduction while providing accurate insight of both augmentation and normalization methods. AugmentHE provided an 81% increase in color dispersion compared to geometric augmentations only. In addition, every classification model that involved AugmentHE presented a significant increase in the area under receiving operator characteristic curve (AUC) over the widely used RGB shift. More precisely, AugmentHE-based models showed at least 0.14 AUC increase over RGB shift-based models. Regarding normalization, HEnorm appeared to be up to 78x faster than conventional methods. It also provided satisfying results in terms of bias reduction. Altogether, our pipeline composed of AugmentHE and HEnorm improved AUC on biased data by up to 21.7% compared to usual augmentations. Conventional normalization methods coupled with AugmentHE yielded similar results while being much slower. In conclusion, we have validated an open-source tool that can be used in any deep learning-based digital pathology project on H&E whole slide images (WSI) that efficiently reduces stain-induced bias and later on might help increase pathologists' confidence when using AI-based products.

* Corresponding author at: Institut Universitaire du Cancer-Oncopole, Pathology Department, F-31059 Toulouse, France.
E-mail addresses: franchet.camille@iuct-oncopole.fr (C. Franchet), brousset.pierre@iuct-oncopole.fr (P. Brousset).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2024.108130
Received 24 October 2023; Received in revised form 7 February 2024; Accepted 7 February 2024
Available online 8 February 2024

1. Introduction

Microscopic analysis relies on the observation of tissue slices using a brightfield microscope. In order to highlight biological structures, these slices are stained using hematoxylin and eosin (H&E). Actually, safran is also used in some laboratories. Hematoxylin stains cell nuclei in blue while eosin highlights cytoplasms and connective tissues in pink. Despite the use of the same chemical compounds, the actual slice colors strongly vary from one laboratory to another. Both dye manufacturers and particular staining protocols from different laboratories cause H&E color variability. Consequently, the origin of a slide can often be inferred from its stain characteristics. Besides, color intensity is altered by time, making old samples brighter and less contrasted than more recent ones. Altogether, these variations represent a frequent source of bias in digital pathology.

Machine Learning (ML) algorithms rely on the assumption that the training dataset accurately represents the diversity of real-world data. Unfortunately, it is hardly ever the case, especially in healthcare datasets in which bias is often induced by overrepresentation of frequent cases, a limited diversity of data sources and the small size of datasets. Such drawbacks can trigger overfitting, giving rise to ML models that are unable to generalize to real-world data. It is worth mentioning that in clinical research, patients' cohorts are statistically correlated to prognostic features on the basis of both inclusion and exclusion criteria. In short, the selection of patients can produce a major bias when building ML models.

A common workaround is to use data augmentation, which consists in modifying images using either geometric (rotation, translation, crop) or pixel level (color, contrast, luminosity, noise) transformations. Pixel-level transformations acting on images hue are commonly used to make model predictions less reliant on stain [1]. Unfortunately, these transformations often generate unrealistic output, creating noisy data that cannot exist in real world. Although little noise is welcome in ML as it helps models to better generalize, excessive noise can hinder their performance. Some attempts have been proposed to reduce stain-induced bias using stain deconvolution [2] to emulate real-world color variability [3] and create realistic images. Beside data augmentation, another procedure to reduce this bias is color normalization. While data augmentation artificially expands a dataset during training, normalization rather reduces variability at inference time. More specifically, normalization modifies colors of input images to make its stain similar to that of a reference dataset or image. This method can be used at training – combined with data augmentation – and at inference time. Even though it seems counterintuitive to concurrently use augmentation and normalization, the latter allows data to be concentrated around a reference point before being spread using augmentation. This enables normalized images to be integrated into the training color space at inference time. Several approaches to normalize images have been already reported. Most relied on the estimation of a color deconvolution matrix that sets a link between H&E stain and RGB space [2,4–6]. More recent methods used ML to identify morphological structures and associate them with one specific stain [7–9]. In addition, some authors used variational autoencoders [10] or generative adversarial networks (GAN) [11] to generate images, its colors matched with a reference stain [12–18]. Although these methods have been widely used in digital pathology, only a few studies described their actual performance regarding bias reduction [19,20].

In our study, we have developed innovative methods for stain augmentation (AugmentHE) and fast normalization (HEnorm), that revisited Vahadane's method [6] to separate hematoxylin and eosin channels so as to better simulate real-world data. The impact of these methods on bias reduction has been studied and compared to other widespread techniques, therefore we used a multicenter dataset created for breast cancer histological grade assessment. Classification models were trained with single cohort images before assessing its performance on images coming from other cohorts. This setting led to extensively monitor bias reduction, thus providing accurate assessment of our augmentation and normalization methods.

2. Materials and methods

2.1. Cohort

A representative subset of available H&E slides from patients included in the PACS04 [21], PACS05 [22] and PACS08 [23] UNICANCER trials were used in our study. PACS04 was a phase III open-label randomized and multicentric study evaluating: (i) combined administration of docetaxel and epirubicin versus fluorouracil, epirubicin and cyclophosphamide (FEC)-100 in the treatment of non-metastatic lymph node-positive breast cancer patients; (ii) sequential addition of trastuzumab for [HER2 3+] or [HER2 2+ and FISH+] patients. PACS05 was a phase III randomized study of 4 versus 6 cycles of adjuvant FEC, in women with stage I breast cancer. PACS08 was a randomized open label multicentric phase III trial that evaluate the benefit of a sequential regimen associating FEC100 and Ixabepilone in adjuvant treatment of non metastatic poor prognosis breast cancer defined as triple negative tumor [HER2 negative – ER negative – PR negative] or [HER2 negative and PR negative] tumor in node positive or node negative patients. A representative surgical specimen paraffin block of the tumor was selected for each patient. H&E slide staining was centrally performed in a distinct laboratory for each trial, i.e. Institut Gustave Roussy (Villejuif, France), Centre Jean Perrin (Clermont-Ferrand, France) and Institut Claudius Regaud (Toulouse, France) for PACS04, PACS05 and PACS08 respectively. Histological grade was based on the evaluation of 3 items : tubule formation, nuclear pleomorphism and mitotic count as described by C.W. Elston and I.O. Ellis in 1991 [24]. Each item is scored between 1 and 3. The sum of the 3 scores defines the histological grade: 3, 4 or 5 = grade I; 6 or 7 = grade II; 8 or 9 = grade III. These three trials were approved by regulatory and ethics authorities of all participating centers (ClinicalTrials.gov identifier, NCT00054587, NCT00055679 and NCT00630032 for PACS04, PACS05 and PACS08, respectively). The project complies with the reference methodology MR-004 and has been registered in the public directory of the Health Data Hub (F20220930101746). The methods were performed in accordance with relevant guidelines and regulations and approved by CNIL and the Projects Review and Ethical Committee of the Institut Claudius Regaud - IUCT-Oncopole. Owing to the retrospective nature of this study and to the respect of the reference methodology MR-004, the ethics committee granted a waiver of informed consent for the included participants.

2.2. Whole Slide Images (WSI) acquisition

H&E slides were digitized using a Pannoramic 250 Flash II digital microscope (3DHISTECH, Budapest, Hungary) at 20× magnification to achieve a scan resolution of 0.24 μm/pixel.

2.3. AugmentHE: Histology-specific method for stain color augmentation

AugmentHE was specifically designed for stain color augmentation. It relies on the assumption that stain color variability of H&E slides is related to (i) the dye composition might differ from one supplier to another. It also leads to significant differences in stain. For example, eosin hue ranges from intense pink to light orange. In addition, some laboratories use safran adding a yellowish hue of collagen bundles; (ii) the slide preparation protocol depends on slice thickness or dye amount, it might also cause variability in luminosity, contrast or color saturation; (iii) the slides oldness when stains tend to become less contrasted and saturated through time. This fading effect can be troublesome for retrospective cohorts with several years follow-up. Therefore, our augmentation method should affect two variables: color produced by both hematoxylin and eosin and color concentration that might be altered by tissue thickness or dye amount. Using Vahadane's method [6] thanks to StainTools library [25], given $I \in \mathbb{R}^{3\times hw}$ (where h and w are the image height and width respectively) the image in

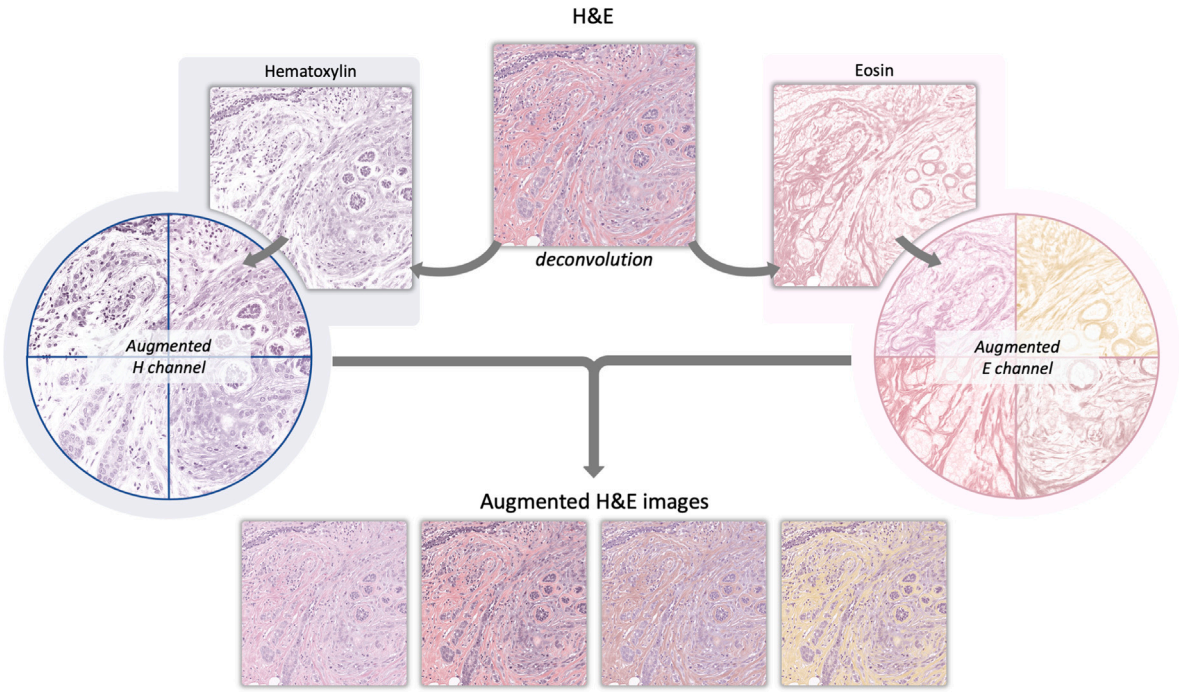


Fig. 1. Illustration of AugmentHE method. Hematoxylin and eosin channels are first extracted from the input image. They are then randomly modified and recombined to create a restained image. Four examples of potential output images are displayed herein.

RGB space, a stain color matrix $W \in \mathbb{R}^{3\times2}$ and a stain density matrix $H \in \mathbb{R}^{2\times hw}$ were computed such that:

$$I = \exp(-WH) \tag{1}$$

The W matrix columns represent colors in RGB space that pure hematoxylin and eosin produce. The H matrix represents the amount of hematoxylin and eosin each pixel contains. This implies that modifying W artificially changes the dye composition, while modifying H changes dye concentration. We therefore sampled $\alpha_W \in [1-\alpha_W^{lim}, 1+\alpha_W^{lim}]^{3\times2}$, $\beta_W \in [-\beta_W^{lim}, \beta_W^{lim}]^{3\times2}$, $\alpha_H \in [1-\alpha_H^{lim}, 1+\alpha_H^{lim}]^2$, $\beta_H \in [-\beta_H^{lim}, \beta_H^{lim}]^2$ (with $\alpha_W^{lim}, \beta_W^{lim}, \alpha_H^{lim}, \beta_H^{lim} \in \mathbb{R}$) from uniform distributions. We thus defined $W' \in \mathbb{R}^{3\times2}$ and $H' \in \mathbb{R}^{2\times hw}$ such that:

$W' = \alpha_W \cdot W + \beta_W \mathbb{I}_{3,2}$, with $(\cdot)$ denoting element-wise product.

$$H' = \begin{cases} \alpha_H H + \beta_H \mathbb{I}_{2,hw} & \text{where } H > k \\ H & \text{otherwise} \end{cases} \tag{2}$$

Using a threshold value k for concentration levels ensures that background remains unchanged. We set $k = 0.2$ in all experiments. We then defined a new altered image $I' \in \mathbb{R}^{3\times hw}$ as:

$$I' = \exp(-W'H') \tag{3}$$

I' is identical to I except for its randomly generated stain. This enabled us to represent a very large diversity of color stains from a single source, which helped us to reduce the associated bias. The whole process is illustrated in Fig. 1.

2.4. HEnorm: deep learning-based method for H&E images normalization

2.4.1. Normacolor dataset

The Normacolor dataset is a 700 fully-anonymized WSI collection of breast tissue slices gathered in the Pathology Department of the University Cancer Institute of Toulouse - Oncopole. This WSI acquisition was done in a short period of time in order to avoid any substantial change in specimen handling, slicing technique and staining protocol. Therefore, this collection was composed of homogeneously stained WSI breast tissue. These WSI were then divided into contiguous 1024 × 1024-pixel patches filtered to remove background, creating more than 64,000 images. This dataset was then split into training and validation sets. Tissue samples were collected and processed at the "CRB Cancer des Hôpitaux de Toulouse", following ethical procedures (Declaration of Helsinki). In accordance with French law, the "CRB Cancer des Hôpitaux de Toulouse" Cancer collection was reported to the Ministry of Higher Education and Research (DC 2009–989). A transfer agreement (AC-2008-820) was obtained following approval by the ethics committees.

2.4.2. Model

HEnorm is a deep learning (DL)-based normalization method that is trained using Normacolor dataset and AugmentHE. At training time, images were randomly re-stained using AugmentHE before being transferred into H&E space using a scikit-image [26] implementation that we adapted for PyTorch [27]. The H channel has been used alone as input to a U-Net model [28] expected to reconstruct the original non-augmented H&E image, since the most relevant features of tissue architecture lie in this channel. As a matter of fact, hematoxylin counterstaining is widely used in routine practice since it allows an accurate appreciation of tissue architecture mainly in tumoral tissues in which tumor cellularity allows pathologists to capture these patterns. Nevertheless, there might be some exceptions such as collagen-rich tissues or some parasitic infections in which key image patterns lies in the eosin channel [29]. Our method forced a full E channel restaining that often exhibits most of the stain variability. The whole model is illustrated in Fig. 2.

We have been using both PyTorch [27] and Pytorch Lightning (https://www.pytorchlightning.ai/) libraries. FastAI's [30] DynamicUnet (https://docs.fast.ai/vision.models.unet.html) has been implemented with custom CGR_5_32_4 encoder (Supplementary Figure 1). Four-image batches have been applied and backpropagation has been performed using Mean Squared Error (MSE) loss with AdamW optimizer [31]. The training process practically followed two steps:

1. 10 epochs with cosine-annealed OneCycle scheduler [32] with a maximum learning rate of 2×10^{-4} and a weight decay of 10^{-3};

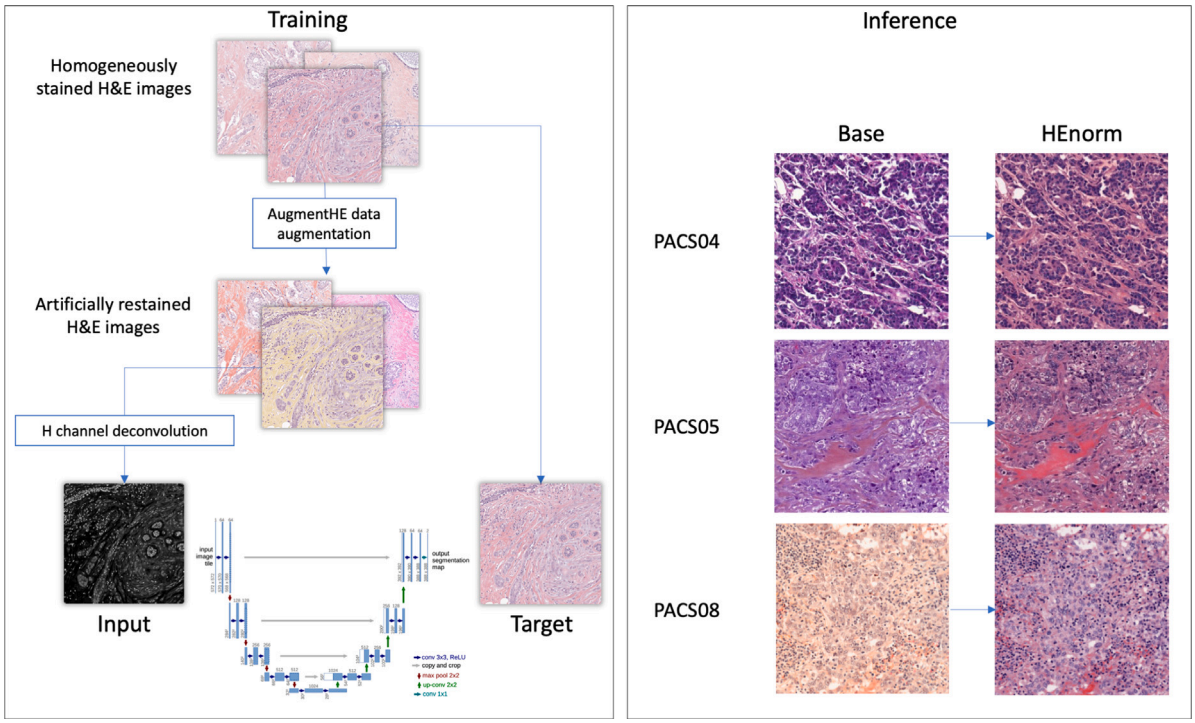


Fig. 2. Illustration of HEnorm method. Images are first augmented using AugmentHE and their hematoxylin (H) channel are extracted. These H channels then went through a U-Net model in order to reconstruct the original non-augmented images. Some results are included on the right.

2. 150 epochs with reduced learning rate on plateau scheduler with early stopping, an initial learning rate of 10^{-3} and a weight decay of 2×10^{-4}.

The total amount of training time was about 600 h using a single Tesla P100 PCIe 16 GB GPU.

2.5. Effectiveness measurement of AugmentHE and HENorm on bias reduction

In order to measure the effectiveness of AugmentHE and HEnorm on bias reduction we focused on a single task: histological grade classification of breast cancer. We chose to focus on grade I and grade III distinction since we have essentially wanted to monitor bias reduction on a straightforward and intuitively simple classification task. The classification of grade I versus grade III BC has already been clinically relevant [33]. Particularly, our goal was to discriminate between histological grades I and III on the basis of H&E WSI. Our aim was also to test whether AugmentHE and HEnorm improved the generalization ability of a model trained on biased data, instead of reaching high accuracy since automated grade assessment remains challenging.

2.5.1. Data

To that end, our dataset was split up depending on the source of the images cohort. As it contained most of these images (Supplementary Table 1), PACS04 was used for training, whereas PACS05 and PACS08 were kept for biased test. Ten percent of PACS04 images were put aside for validation while the remaining 90% were divided into 5 folds for k-fold cross-validation. Tissue detection was then computed using thresholding in the L*a*b* space on WSI thumbnails, before extracting contiguous 1024 × 1024-pixel patches in foreground regions. This resulted in about 420,000 PACS04 patches, being 80,000 histological grade I and 340,000 histological grade III respectively. In order to face this data imbalance, we randomly undersampled grade III images during training step, showing up a difference of less than 3% in class populations. This is to say that only 40,000 images were used at each epoch. During the inference step, 20,000 PACS05 patches (3000 grade I and 17,000 grade III) and 28,300 PACS08 patches (800 grade I and 27,500 grade III) were applied.

2.5.2. Model

We tried different combinations of bias reducing techniques, all including at least geometric augmentations, in order to compare them between each other. Our baseline used only geometric augmentations such as flips and transpositions (noted as basic). We compared it to a classical augmentation method called RGB shift which applies a random offset to the values of each RGB channel (noted as aug_classic). We then evaluated both AugmentHE and HEnorm separately (respectively noted as augHE and HEnorm) and both combined (noted as HEnorm_augHE). In order to compare HEnorm to the state of the art, we evaluated combined AugmentHE and Vahadane's normalization method [6] (noted as Vahadane_augHE). For each of these combinations, we trained 10 ResNet50 [34] classifiers, using two series of 5-fold cross-validation. Backpropagation was performed using binary cross-entropy loss and AdamW optimizer [31]. We used a One Cycle [32] policy with a maximum learning rate of 2×10^{-4} and a weight decay of 10^{-3}. Training was performed using batches of 8 images for 10 epochs, with a single Tesla P100 PCIe 16 GB GPU.

3. Data availability

The WSI from the three cohorts PACS04, PACS05 and PACS08 are the property of Unicancer. Unicancer certifies that it will share de-identified individual data that underlie the results reported under the following conditions: - Unicancer will grant access to all study data to all requestors in the field of cancer research, upon written detailed request sent to DAC-OS2006@unicancer.fr - Unicancer will grant access to editor's reviewers to study documents and to the data required for independent verification of the published results. Editor guarantees that mandated reviewers are committed to use the data transferred for the sole aim to reviewing produced results, under strict conditions warranting security and confidentiality of the data.

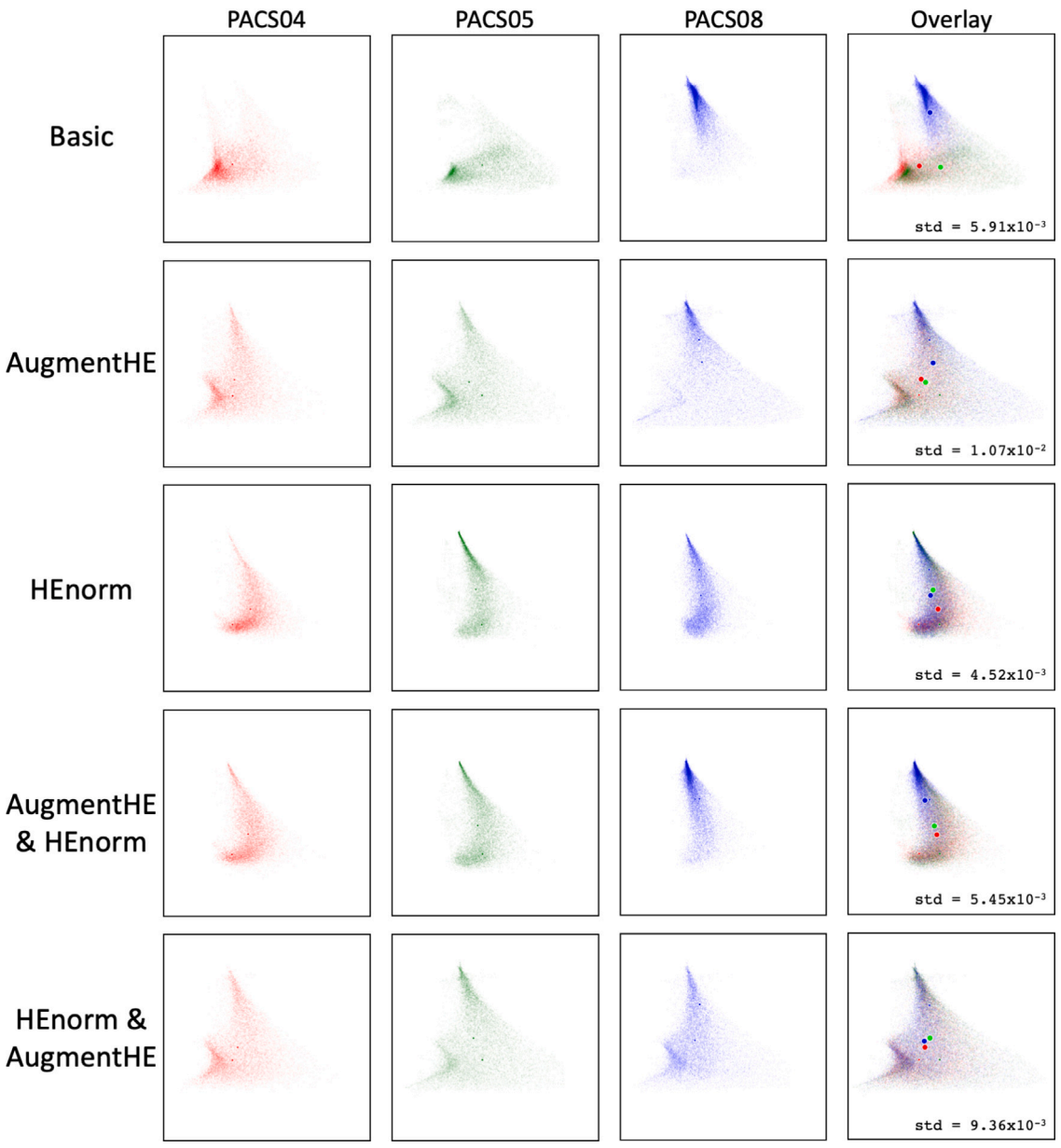


Fig. 3. 2D representation of color dispersion when using different sets of transformations. Points are 2D orthogonal projections of normalized hue histograms from the different cohorts. The projection plane contains the centroids from the three cohorts with no transformation applied. These centroids are highlighted on the overlay column (for comparison, centroids from the basic setup are also highlighted). Average standard deviation across all histogram bins is also given.

4. Code availability

Source code for HEnorm and AugmentHE is available at https://github.com/schwobr/HEnorm. It includes pretrained weights for HEnorm on 1024 × 1024 images at level 1 (0.5 μm/px).

5. Results

5.1. PACS04, PACS05 and PACS08 WSI exhibited a major color bias

On the one hand, the cohorts had been previously stained in distinct central laboratories, resulting in a color palette belonging to each of them. On the other hand, inclusion criteria of these clinical trials were disparate. Indeed, PACS04 and PACS05 patients harbored rather good prognoses whereas PACS08 patients showed less favorable outcomes. Consequently, a direct link between color palette and aggressiveness of tumors appeared. To address the question whether both data augmentation and color normalization might reduce this bias, we have randomly selected a subset of grade I and grade III tumors following the overall cohort proportions. Thus, we have included 128 grade I and 436 grade III tumors from PACS04, 9 grade I and 221 grade III tumors from PACS05, 3 grade I and 328 grade III tumors from PACS08 respectively.

5.2. AugmentHE pragmatically increased dataset color dispersion

The purpose of color augmentation is to better cover the variability of real data. Basically, regarding microscopic images, a special attention must have been paid to preserve biological information. AugmentHE achieved this goal by generating credible pathological images (Fig. 1) while color dispersion successfully increased. Indeed, we assessed the normalized hue histograms of all images and computed their standard deviation for each bin. The average standard deviation (ASD) increased from 5.91×10^{-3} with only geometric augmentations to 1.07×10^{-2} with AugmentHE (Fig. 3), providing a 81% rise. We included the 2D projections of these hue histograms on a plane defined by the centroids

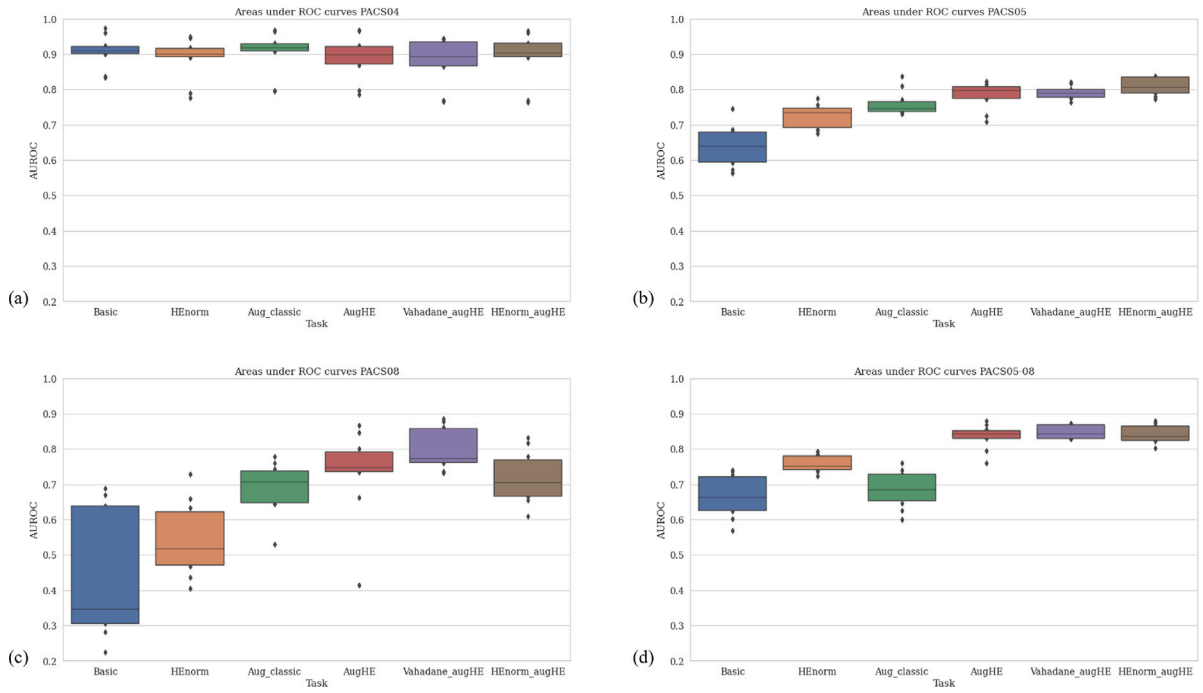


Fig. 4. The areas under receiver operating characteristic curves (AUC) from 10 models trained on PACS04 images for each task on (a) PACS04 validation set (b) PACS05 whole dataset (c) PACS08 whole dataset and (d) combined PACS05 and PACS08 datasets. Augmentation settings are labeled as follows: Basic = only geometric augmentations (rotation, flip, transpose); AugHE = Basic + AugmentHE; Aug_classic = Basic + RGB shift. In addition, both HEnorm and Vahadane's methods [6] were combined with AugmentHE. All average AUC values can be found in Supplementary Table 7.

of the 3 cohorts (with no augmentation applied), in Fig. 3. These point clouds clearly confirmed the increase in data dispersion when using AugmentHE. Besides, the centroids of the cohorts were brought closer together by these augmentations, which proved that AugmentHE made the cohorts color distributions harder to distinguish.

5.3. HEnorm efficiently normalized stain while preserving image structure

HEnorm managed to normalize colors while preserving tissue structure at a low computational cost. To compare, it took 30 ms to normalize a 1024 × 1024-pixel image using HEnorm, being 3.5× faster than Reinhard's [4], 50× faster than Macenko's [5] and 78× faster than Vahadane's [6] methods respectively (Supplementary Table 2). Despite HEnorm presented a slightly less precise image structure preservation than the previously mentioned methods, the structural similarity (SSIM) and Pearson correlation coefficient (PCC), ranging from 0 to 1, attained 0.891 and 0.889 respectively (Supplementary Table 2).

Concerning reduction of differences in color distribution between cohorts, Fig. 3 shows that HEnorm decreased the ASD of normalized hue histograms by 25%. Furthermore, point clouds are visually overlapping, which is illustrated by bringing closer centroids of these 3 cohorts.

5.4. Combination of both AugmentHE and HEnorm contributed to bias reduction

When it comes to ML algorithm training, cohorts containing images from different origins should be indistinguishable in order to avoid any bias. To that end, it is common to chain normalization and augmentation, also leading to wide enough data dispersion so as to prevent overfitting. Intuitively, applying these transformations in reverse order would be counterproductive as normalization should cancel out stain augmentation. This is confirmed in Fig. 3, where point clouds obtained when using AugmentHE before HEnorm looked alike those obtained using HEnorm only. Furthermore, ASD only increased by 21% when adding AugmentHE before HEnorm. In contrast, when using AugmentHE only, ASD was 137% higher than with HEnorm. Conversely, adding HEnorm before AugmentHE only decreased ASD by 12%, with the added benefit of a near perfect overlap between point clouds and its centroids. Consequently, the use of HEnorm followed by AugmentHE looked like the optimal combination for training step. Practically, the main goal of our study was to monitor generalizability of models trained on biased datasets applying different combinations of data augmentation and normalization. On the one hand, we determined that augmentation and normalization caused no significant decrease in validation area under receiving operator characteristic curve (AUC) on PACS04 (Supplementary Table 3) regarding AUC values ranging from 0.88 ± 0.06 for Vahadane_augHE to 0.91 ± 0.04 for basic. On the other hand, augmentation and normalization significantly increased AUC on test data from combined PACS05 and PACS08 (Fig. 4 and Supplementary Table 6). More specifically, every combination that involved AugmentHE caused a significant increase in AUC over basic or aug_classic. According to this, AUC ranged from 0.67 ± 0.06 and 0.69 ± 0.05 for basic and aug_classic respectively, and reached up to 0.83 ± 0.03, 0.84 ± 0.02 and 0.85 ± 0.02 for augHE, HEnorm_augHE and Vahadane_augHE respectively. Similarly, only models trained without AugmentHE showed significant decrease in AUC from validation to test steps, with a 25% reduction on average (Supplementary Table 8). More technically, loss function values are presented in Supplementary Figure 2.

Looking at results on PACS05 and PACS08 separately, some questions raised concerning the effectiveness of each combination. In particular, Vahadane_augHE showed no significant increase in AUC over aug_classic or HEnorm on PACS05, whereas HEnorm_augHE had no significant increase in AUC over aug_classic on PACS08 (Supplementary Tables 4 and 5). Strikingly, all methods showed a significant decrease in AUC from validation to test steps on separated PACS05 and PACS08 with the exception of Vahadane_augHE on PACS08 (Supplementary Table 8). A noticeable difference thus appeared between results obtained when combining PACS05 and PACS08 compared to those obtained on them separately. For instance, AUC coming from HEnorm_augHE experiment was 0.85 ± 0.02 on pooled PACS05 and PACS08, whereas

it was only 0.82 ± 0.02 on PACS05 and 0.73 ± 0.07 on PACS08. This difference can be explained by the massive imbalance between grade I and grade III in PACS05 and PACS08. Indeed, even with a model that would classify all PACS08 images as grade III, the false positive rate on combined PACS05 and PACS08 cohorts would remain rather low as grade I tumors of PACS08 only represent 25% of all grade I images. On the contrary, when working on PACS08 alone, predicting all images to be grade III implies having a false positive rate of 100%. It is the same case for PACS05 to a lesser degree, as it contains a higher amount of grade I images than PACS08.

6. Discussion

Two innovative approaches for data augmentation (AugmentHE) and color normalization (HEnorm) have been developed in our study, providing an end-to-end pipeline for histological images preparation within DL projects. Respective impacts of AugmentHE and HEnorm have been assessed and compared to widely used techniques previously described in the literature [4–6]. Thus, the most efficient combination of both AugmentHE and HEnorm regarding bias reduction has been established herein. As pathologists, we developed AugmentHE as a pragmatic color augmentation method based on our daily practice in laboratories. This method takes advantage of our tangible knowledge of the staining process as a broad source of color variability. It was also greatly inspired by Tellez et al. [19] as their HED augmentation helped us to translate our knowledge regarding the staining process into an actual augmentation pipeline. Thanks to it, we generated quite convincing and realistic histological images compared to those produced using color augmentation methods based on non-specific approaches such as RGB or hue shifts. More importantly, a significantly better bias reduction was obtained with AugmentHE compared to classical methods. Furthermore, HEnorm appeared to be a fast and reliable method for color normalization that considered the major role of eosin composition in color variation of H&E slides. Actually, eosin is frequently associated with orange or yellow dyes that drastically modify the hue of H&E images from one laboratory to another. In addition, eosin fades with time and light exposure rather than hematoxylin, leading to increased color variability. On the contrary, hematoxylin is shown to be more stable and harbor most of the characteristics about tissue structure that can be accurately used as a canvas for H&E image reconstruction. Accordingly, HEnorm is a DL-based H&E generator from hematoxylin input images providing normalized output images. In terms of results, HEnorm efficiently decreased color dispersion from one cohort to another and was much faster than every other method herein tested [4–6]. Interestingly, according to generalizability, our method significantly reduced bias when combined with AugmentHE concerning PACS05 images. However, our model behaved differently regarding PACS08 cohort in which Vahadane's method combined with AugmentHE provided the best bias reduction results. Concerning results from color normalization, neither HEnorm nor Vahadane's method alone could fully reach bias reduction expectations. Therefore, it clearly appeared that color normalization could be regarded as a fine tuning tool applied at the end of ML model development to further reinforce generalizability. Besides, results were significantly better on combined PACS05 and PACS08 than on each cohort alone. It is a clear illustration that good results can hide bias if they are not properly analyzed through more precise lenses. We assessed both when and how to apply AugmentHE and HEnorm by exploring every possible combination in order to provide the most optimal sequence in terms of bias mitigation. We therefore showed that the best way to bring the hue of cohorts closer is to firstly apply normalization and then data augmentation, thus it considerably reduced stain-induced bias. Besides these satisfying results, several options should still be considered as potential enhancements of our pipeline. Firstly, AugmentHE might benefit from an even more realistic modelization of color variations between laboratories. In fact, we chose to enlarge color dispersion on the basis of our laboratory staining color space. Another option could have been to directly represent real world color variation for which unfortunately there is still no precise knowledge available. In practice, defining such a color space would require gathering a large multicenter dataset as an acceptable representation of real-world variability. Secondly, in terms of color normalization, Tellez et al. showed that the mix of several normalization methods can help to reduce bias [19]. The latter study corroborates our results showing that there is no universal normalization method that is effective on all datasets. Thirdly, we identified potential improvements on HEnorm method that would rely on adversarial learning techniques. Indeed, HEnorm was trained using a simple MSE loss, which has been perfectly designed for image reconstruction despite not being state-of-the-art. Therefore, this reconstruction may benefit from adversarial loss, as generated images should not be distinguishable from the original ones [13,35]. Actually, such adversarial techniques have already been described in the literature [12–15,18] and demonstrated better results than classical methods did [36]. We admit this is one out our study's main limitations and the impact of those more recent approaches on classifier performances, combined or not with AugmentHE, should be addressed in future works. Another aspect that could be considered as a limitation is the choice to present classification performance only at patch level, without aggregating results at WSI level. Indeed, the spatial heterogeneity of breast cancers may account for variations in grade at patch level that should be considered when aggregating data at WSI level [37]. However, it is important to emphasize that the primary objective of our study was to monitor bias reduction through some normalization and augmentation tools, including ours. Consequently, performances evolution at patch level was sufficient to assess the impact of various combinations of data augmentation and normalization. Furthermore, recent papers have moved away from the concept of grade to directly classify tumors based on the presence or absence of early BC recurrence [38]. However, we speculate that these approaches would also benefit from the use of bias reduction tools, such as those proposed herein. Concerning the fact that our work is fully based on BC samples, both AugmentHE and HEnorm can be applied to other cancer types. On the one hand, AugmentHE is already applicable to every kind of H&E histological images. On the other hand, HEnorm could either work on other types of cancer without being modified, or might need an additional training based on targeted tissue. Moreover, we remain uncertain that HEnorm would be efficient in rare cases in which key image patterns lie in the eosin channel, such as collagen-rich tissues or some parasitic infections. Further research should be carried out in order to validate HEnorm in these cases.

7. Conclusion

Our work is thus part of a general breakpoint in ML history where there is a growing interest in the study of bias after a long period of time where proof of concepts have been at the forefront of ML research. Indeed, we quantified stain-induced bias in a multisite dataset as well as the effectiveness of several augmentation and normalization methods, including our own. According to the fact that nowadays datasets are hardly ever optimally designed for ML studies, we propose a pipeline that helps mitigating bias in ML models for histological images. Our pipeline should also serve as a baseline for future proof of concepts that need a ready-to-use bias-reduction method as well as new studies on bias in digital pathology. Indeed, thanks to its versatility, our method is not only designed for color-bias reduction in breast cancer studies but can also adapt to pan-cancer studies.

CRediT authorship contribution statement

Camille Franchet: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Robin Schwob:** Writing – review & editing, Writing – original draft,

Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Guillaume Bataillon:** Validation. **Charlotte Syrykh:** Validation. **Sarah Péricart:** Validation. **François-Xavier Frenois:** Visualization, Software. **Frédérique Penault-Llorca:** Resources, Funding acquisition. **Magali Lacroix-Triki:** Resources, Funding acquisition. **Laurent Arnould:** Resources, Funding acquisition. **Jérôme Lemonnier:** Resources, Data curation. **Jean-Marc Alliot:** Conceptualization. **Thomas Filleron:** Validation, Methodology. **Pierre Brousset:** Writing – original draft, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper has benefited from the Imag'IN Platform of the Institut Universitaire du Cancer - https://www.imagin-labs.net/.

This paper has also benefited from the proofreading of Ana Darquier, a native specialized medical translator.

This work has been supported by funds from Bpifrance (APRIORICS project), Thales Services Numérique, Fondation pour la Recherche Médicale, France, Agence Nationale de la Recherche, France, Institut Carnot CALYM (DIAL project) and Laboratoire d'Excellence Toulouse Cancer (TouCan). This work has also benefited from technical support by Health Data Hub (APRIORICS project).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108130.

References

[1] Y. Liu, et al., Detecting cancer metastases on gigapixel pathology images, 2017, arXiv:1703.02442 [cs], URL http://arxiv.org/abs/1703.02442.

[2] A. Ruifrok, D. Johnston, Quantification of histochemical staining by color deconvolution, Anal. Quant. Cytol. Histol. 23 (2001).

[3] D. Tellez, et al., Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks, IEEE Trans. Med. Imaging 37 (9) (2018) 2126–2136, URL http://arxiv.org/abs/1808.05896.

[4] E. Reinhard, M. Ashikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. (2001) 8.

[5] M. Macenko, et al., A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE International Symposium on Biomedical Imaging: From Nano To Macro, IEEE, Boston, MA, USA, 2009, pp. 1107–1110, URL http://ieeexplore.ieee.org/document/5193250/.

[6] A. Vahadane, et al., Structure-preserving color normalization and sparse stain separation for histological images, IEEE Trans. Med. Imaging 35 (8) (2016) 1962–1971.

[7] A. Khan, N. Rajpoot, D. Treanor, D. Magee, A non-linear mapping approach to stain normalisation in digital histopathology images using image-specific colour deconvolution, IEEE Trans. Biomed. Eng. 61 (2014).

[8] B.E. Bejnordi, et al., Stain specific standardization of whole-slide histopathological images, IEEE Trans. Med. Imaging 35 (2) (2016) 404–415.

[9] D. Bug, et al., Context-based normalization of histological stains using deep convolutional features, in: M.J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J.M.R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J.P. Papa, A. Madabhushi, J.C. Nascimento, J.S. Cardoso, V. Belagiannis, Z. Lu (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, in: Lecture Notes in Computer Science, Springer International Publishing, 2017, pp. 135–142.

[10] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014, arXiv:http://arxiv.org/abs/1312.6114v10.

[11] I. Goodfellow, et al., Generative adversarial nets, in: Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014, URL https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[12] A. Janowczyk, A. Basavanhally, A. Madabhushi, Stain normalization using sparse AutoEncoders (StaNoSA): Application to digital pathology, Comput. Med. Imaging Graph. 57 (2017) 50–61, URL https://linkinghub.elsevier.com/retrieve/pii/S0895611116300404.

[13] H. Cho, S. Lim, G. Choi, H. Min, Neural stain-style transfer learning using GAN for histopathological images, 2017, arXiv:1710.08543 [cs]. URL http://arxiv.org/abs/1710.08543.

[14] E. Yuan, J. Suh, Neural stain normalization and unsupervised classification of cell nuclei in histopathological breast cancer images, 2018, arXiv:1811.03815 [cs, q-bio] URL http://arxiv.org/abs/1811.03815. arXiv:1811.03815.

[15] F.G. Zanjani, S. Zinger, B.E. Bejnordi, J.A.W.M. van der Laak, P.H.N. de With, Stain normalization of histopathology images using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 573–577, ISSN: 1945-8452.

[16] M.T. Shaban, C. Baur, N. Navab, S. Albarqouni, Staingan: Stain style transfer for digital histological images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, 2019, pp. 953–956.

[17] H. Kang, et al., StainNet: A fast and robust stain normalization network, Front. Med. 8 (2021) URL https://www.frontiersin.org/articles/10.3389/fmed.2021.746307.

[18] N. Bouteldja, et al., Tackling stain variability using CycleGAN-based stain augmentation, J. Pathol. Inform. 13 (2022) 100140, URL https://www.sciencedirect.com/science/article/pii/S2153353922007349.

[19] D. Tellez, et al., Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology, Med. Image Anal. 58 (2019) 101544, URL http://arxiv.org/abs/1902.06543.

[20] J. Boschman, et al., The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images, J. Pathol. 256 (1) (2022) 15–24, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/path.5797. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5797.

[21] M. Spielmann, et al., Trastuzumab for patients with axillary-node–positive breast cancer: Results of the FNCLCC-PACS 04 trial, J. Clin. Oncol. 27 (36) (2009) 6129–6134, http://dx.doi.org/10.1200/JCO.2009.23.0946, PMID: 19917839.

[22] P. Kerbrat, et al., Optimal duration of adjuvant chemotherapy for high-risk node-negative (N–) breast cancer patients: 6-year results of the prospective randomised multicentre phase III UNICANCER-PACS 05 trial (UCBG-0106), Eur. J. Cancer 79 (2017) 166–175, URL https://www.sciencedirect.com/science/article/pii/S0959804917308225.

[23] M. Campone, et al., UCBG 2-08: 5-year efficacy results from the UNICANCER-PACS08 randomised phase III trial of adjuvant treatment with FEC100 and then either docetaxel or ixabepilone in patients with early-stage, poor prognosis breast cancer, Eur. J. Cancer 103 (2018) 184–194, URL https://www.sciencedirect.com/science/article/pii/S0959804918309390.

[24] C. Elston, I. Ellis, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up, Histopathology 19 (5) (1991) 403–410, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2559.1991.tb00229.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2559.1991.tb00229.x.

[25] P. Byfield, StainTools, a package for tissue image stain normalization, augmentation and more, 2018, https://github.com/Peter554/StainTools.

[26] S. van der Walt, et al., Scikit-image: Image processing in Python, PeerJ 2 (2014) e453, http://dx.doi.org/10.7717/peerj.453.

[27] A. Paszke, et al., Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[28] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, CoRR abs/1505.04597, URL http://arxiv.org/abs/1505.04597.

[29] J.W. Goding, Immunohistology, in: J. Fagerberg, D.C. Mowery, R.R. Nelson (Eds.), Monoclonal Antibodies (Third Edition), Elsevier, 1996, pp. 400–423, Ch. 13.

[30] J. Howard, S. Gugger, Fastai: A layered API for deep learning, 2020, CoRR abs/2002.04688, URL https://arxiv.org/abs/2002.04688.

[31] I. Loshchilov, F. Hutter, Fixing weight decay regularization in Adam, 2017, CoRR abs/1711.05101, URL http://arxiv.org/abs/1711.05101.

[32] L.N. Smith, N. Topin, Super-convergence: Very fast training of residual networks using large learning rates, 2017, CoRR abs/1708.07120. URL http://arxiv.org/abs/1708.07120.

[33] Y. Wang, et al., Improved breast cancer histological grading using deep learning, Ann. Oncol. 33 (1) (2022) 89–98, URL https://www.sciencedirect.com/science/article/pii/S0923753421044860.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, CoRR abs/1512.03385. URL http://arxiv.org/abs/1512.03385.

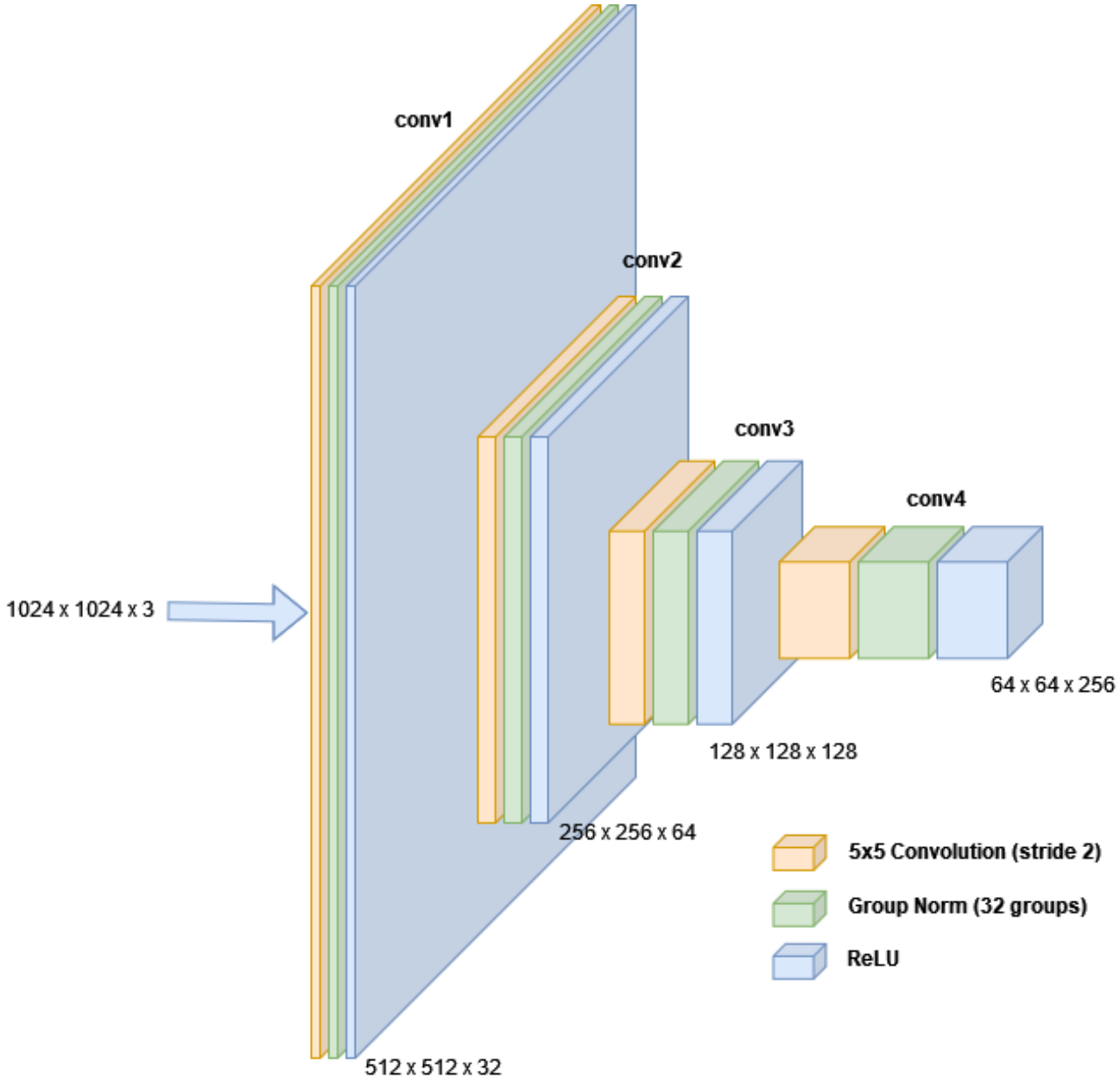
[35] A.-K. Thebille, et al., Deep learning-based bias transfer for overcoming laboratory differences of microscopic images, in: B.W. Papież, M. Yaqub, J. Jiao, A.I.L. Namburete, J.A. Noble (Eds.), Medical Image Understanding and Analysis, in: Lecture Notes in Computer Science, Springer International Publishing, 2021, pp. 322–336.

[36] W. Voon, et al., Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images, Sci. Rep. 13 (1) (2023) 20518, URL https://www.nature.com/articles/s41598-023-46619-6. Number: 1 Publisher: Nature Publishing Group.

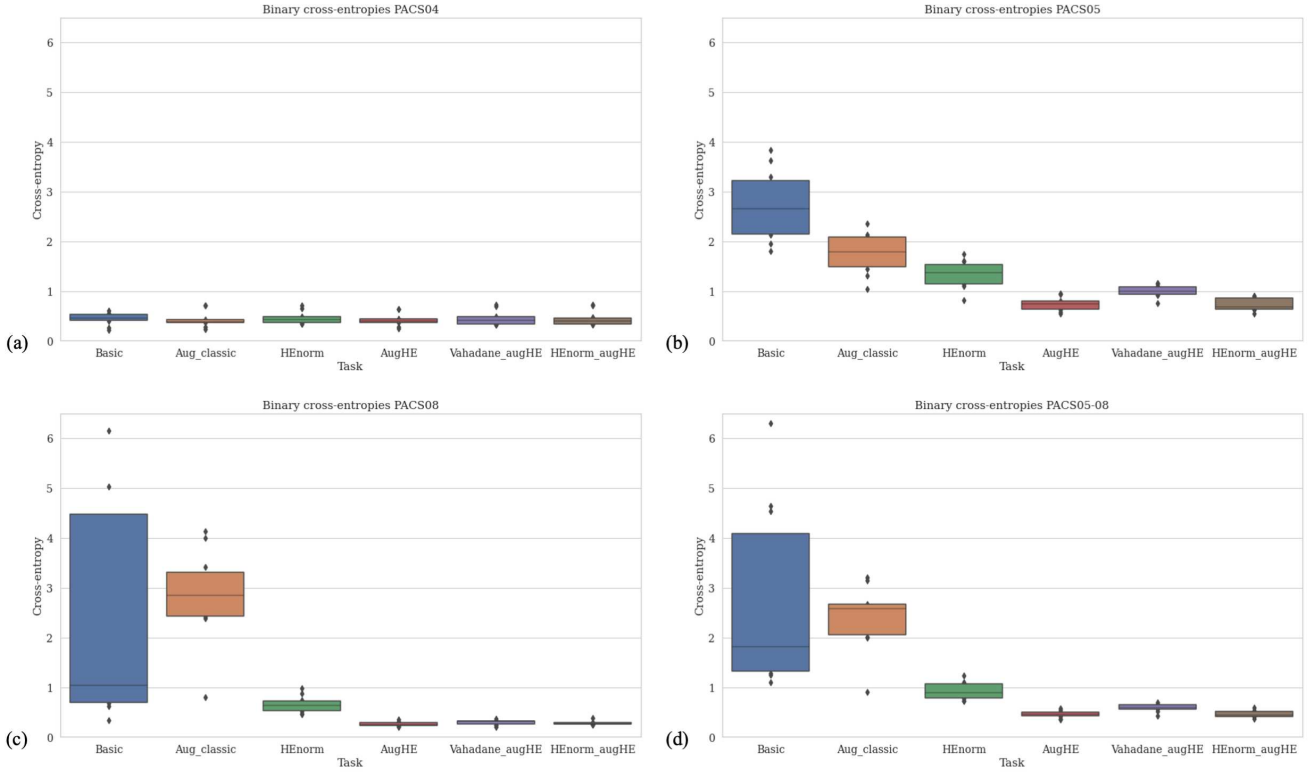
[37] N. Wahab, et al., AI-enabled routine H&E image based prognostic marker for early-stage luminal breast cancer, npj Precis. Oncol. 7 (1) (2023) 1–13, URL https://www.nature.com/articles/s41698-023-00472-y. Number: 1 Publisher: Nature Publishing Group.

[38] Y. Shi, et al., Predicting early breast cancer recurrence from histopathological images in the Carolina Breast Cancer Study, npj Breast Cancer 9 (1) (2023) 1–7, URL https://www.nature.com/articles/s41523-023-00597-0. Number: 1 Publisher: Nature Publishing Group.

Supplementary data



Supplementary Figure 1. CGR_k_w_d designates a neural network composed of d successive layers of $k \times k$ convolutions of stride 2 and width $w \times 2^n$, n being the depth of the current layer (counting from 0), followed by GroupNorm[39] and ReLU. This figure is an illustration of CGR_5_32_4, which was used as an encoder for HEnorm.



Supplementary Figure 2. The binary cross-entropies from 10 models trained on PACS04 images for each task on a) PACS04 validation set, b) PACS05 whole dataset, c) PACS08 whole dataset and d) combined PACS05 and PACS08 datasets. Augmentation settings are labeled as follows: Basic = only geometric augmentations (rotation, flip, transpose); AugHE = Basic + AugmentHE; Aug_classic = Basic + RGB shift. In addition, both HEnorm and Vahadane's method[6] were combined with AugmentHE.

Feature	Category	Not included	Included	p value
Cohort	PACS04	564 (50.1%)	974 (61.5%)	
	PACS05	230 (20.4%)	340 (21.5%)	
	PACS08	331 (29.4%)	269 (17.0%)	
Patients age	<35	55 (3.5%)	49 (4.4%)	0.4831
	35-50	710 (44.9%)	497 (44.2%)	
	>50	818 (51.7%)	579 (51.5%)	
	Median	51	51	
	Range	[22; 71]	[25; 71]	
Surgery type	Conservative	1165 (73.6%)	824 (73.2%)	0.8389
	Mastectomy	418 (26.4%)	301 (26.8%)	
Histological size (mm)	≤20	902 (57.2%)	493 (43.9%)	
	]20; 50]	612 (38.8%)	581 (51.7%)	
	>50	62 (3.9%)	50 (4.4%)	
	Missing	7	1	
Histological grade	Non gradable	2 (0.1%)	0 (0%)	
	I	75 (4.8%)	140 (12.4%)	
	II	991 (62.9%)	0 (0%)	
	III	508 (32.2%)	985 (87.6%)	
	Missing	7	0	
Lymphovascular invasion	No	826 (56.8%)	587 (56.1%)	
	Yes	628 (43.2%)	460 (43.9%)	
	Missing	129	78	
Invaded nodes	0	430 (27.2%)	379 (33.7%)	
	1-3	818 (51.7%)	520 (46.3%)	
	>3	335 (21.2%)	225 (20%)	
	Missing	0	1	
ER	Negative	455 (28.7%)	581 (51.6%)	
	Positive	1128 (71.3%)	544 (48.4%)	
	Missing	0	0	
PR	Negative	650 (43.5%)	706 (66.3%)	
	Positive	844 (56.5%)	359 (33.7%)	
	Missing	89	60	
HER2	Negative	1328 (86.3%)	838 (79.4%)	
	Positive	210 (13.7%)	218 (20.6%)	
	Missing	45	69	
PACS04 immunophenotype	Triple negative	87 (15.5%)	80 (8.2%)	
	HR+/HER2-	287 (51.1%)	694 (71.3%)	
	HER2+	188 (33.5%)	199 (20.5%)	
	Missing	2	1	
PACS05 immunophenotype	Triple negative	63 (39.1%)	52 (17.6%)	
	HR+/HER2-	68 (42.2%)	232 (78.6%)	
	HER2+	30 (18.6%)	11 (3.7%)	
	Missing	69	45	
PACS08 immunophenotype	Triple negative	280 (84.6%)	195 (72.5%)	
	HR+/HER2-	51 (15.4%)	74 (27.5%)	
	HER2+	0	0	
	Missing	0	0	

Supplementary Table 1. Clinical information for the whole cohort. Only part of the cohort was included in the study. We ensured that included and not included data did not show significant statistical difference in patients age and surgery types (p values are computed using χ^2 test). ER = Estrogen Receptor; HER2 = Human Epidermal growth factor Receptor 2; HR = Hormone Receptors; PR = Progesterone Receptor.

Method	time/image	SSIM	PCC
Reinhard	0.11s	**0.936**	**0.994**
Macenko	1.50s	0.898	0.944
Vahadane	2.33s	0.928	0.968
HEnorm	**0.03s**	0.891	0.889

Supplementary Table 2. Comparison of execution time per image, Structural Similarity (SSIM) and Pearson Correlation Coefficient (PCC) for several normalization methods on our dataset. Best results are highlighted in bold. StainTools[25] implementation of Reinhard's, Macenko's and Vahadane's methods was used.

	Basic	Aug classic	AugHE	HEnorm	Vahad AugHE	HEnorm AugHE
Basic	N.A.	0.846 ↓	0.232 ↑	**0.0273** ↑	0.0645 ↑	0.375 ↑
Aug classic		N.A.	0.16 ↑	**0.0137** ↑	**0.0273** ↑	0.131 ↑
AugHE			N.A.	0.557 ↑	0.16 ↑	0.922 ↓
HEnorm				N.A.	0.232 ↑	0.557 ↓
Vahad AugHE					N.A.	**0.0371** ↓

Supplementary Table 3. p values for the Wilcoxon signed-rank test on PACS04 validation AUCs. The ↑ sign means that the mean AUC from the line method is higher than that from the column method. Methods that are significantly different ($p < 0.05$) are highlighted in bold.

	Basic	Aug_classic	AugHE	HEnorm	Vahad AugHE	HEnorm AugHE
Basic	N.A.	**0.00195** ↓	**0.00195** ↓	**0.00391** ↓	**0.00586** ↓	**0.00195** ↓
Aug classic		N.A.	0.084 ↓	**0.00977** ↑	0.193 ↓	**0.00391** ↓
AugHE			N.A.	**0.00391** ↑	0.77 ↓	**0.00391** ↓
HEnorm				N.A.	0.084 ↓	**0.00195** ↓
Vahad AugHE					N.A.	**0.0273** ↓

Supplementary Table 4. p values for the Wilcoxon signed-rank test on PACS05 AUCs. The ↑ sign means that the mean AUC from the line method is higher than that from the column method. Methods that are significantly different ($p < 0.05$) are highlighted in bold.

	Basic	Aug classic	AugHE	HEnorm	Vahad AugHE	HEnorm AugHE
Basic	N.A.	**0.0195** ↓	**0.0195** ↓	0.275 ↓	**0.00195** ↓	**0.0137** ↓
Aug classic		N.A.	0.131 ↓	**0.00977** ↑	**0.00586** ↓	0.492 ↓
AugHE			N.A.	**0.00391** ↑	0.16 ↓	0.16 ↑
HEnorm				N.A.	**0.00195** ↓	**0.00195** ↓
Vahad AugHE					N.A.	**0.00195** ↑

Supplementary Table 5. p values for the Wilcoxon signed-rank test on PACS08 AUCs. The ↑ sign means that the mean AUC from the line method is higher than that from the column method. Methods that are significantly different ($p < 0.05$) are highlighted in bold.

	Basic	Aug classic	AugHE	HEnorm	Vahad AugHE	HEnorm AugHE
Basic	N.A.	0.557 ↓	**0.00195** ↓	**0.00195** ↓	**0.00195** ↓	**0.00195** ↓
Aug classic		N.A.	**0.00195** ↓	**0.00391** ↓	**0.00195** ↓	**0.00195** ↓
AugHE			N.A.	**0.00195** ↑	0.275 ↓	0.557 ↓
HEnorm				N.A.	**0.00195** ↓	**0.00195** ↓
Vahad AugHE					N.A.	0.322 ↑

Supplementary Table 6. p values for the Wilcoxon signed-rank test on combined PACS05 and PACS08 AUCs. The ↑ sign means that the mean AUC from the line method is higher than that from the column method. Methods that are significantly different ($p < 0.05$) are highlighted in bold.

PACS	**Basic**	**Aug classic**	**AugHE**	**HEnorm**	**Vahad AugHE**	**HEnorm AugHE**
04	**0.91 (0.04)**	0.90 (0.06)	0.89 (0.06)	0.89 (0.06)	0.88 (0.06)	0.89 (0.07)
05	0.64 (0.06)	0.76 (0.03)	0.78 (0.04)	0.72 (0.03)	0.79 (0.02)	**0.81 (0.02)**
08	0.44 (0.18)	0.69 (0.07)	0.73 (0.12)	0.54 (0.10)	**0.80 (0.06)**	0.72 (0.07)
05&08	0.67 (0.06)	0.69 (0.05)	0.83 (0.03)	0.76 (0.02)	**0.85 (0.02)**	0.84 (0.02)

Supplementary Table 7. Mean AUC (with standard deviation) on validation (PACS04) and test (PACS05, PACS08 and both combined) sets. Highest value for each dataset in highlighted in bold.

PACS		**Basic**	**Aug classic**	**AugHE**	**HEnorm**	**Vahad AugHE**	**HEnorm AugHE**
05	ΔAUC	0.27 (0.09)	0.15 (0.09)	0.11 (0.08)	0.16 (0.08)	0.11 (0.07)	0.082 (0.085)
	p value	**0.00195**	**0.00977**	**0.00977**	**0.00195**	**0.00391**	**0.0371**
08	ΔAUC	0.47 (0.18)	0.22 (0.09)	0.15 (0.17)	0.35 (0.14)	0.081 (0.117)	0.18 (0.13)
	p value	**0.00195**	**0.00195**	**0.0273**	**0.00195**	0.0645	**0.00977**
05&08	ΔAUC	0.24 (0.09)	0.22 (0.09)	0.057 (0.081)	0.13 (0.071)	0.034 (0.079)	0.049 (0.088)
	p value	**0.00195**	**0.00195**	0.105	**0.00977**	0.232	0.131

Supplementary Table 8. Mean differences in AUC (with standard deviation) between validation (on PACS04) and test (on PACS05 and PACS08) with p values for the Wilcoxon signed-rank test. Significant p values ($p < 0.05$) are highlighted in bold.

III. DISCUSSION

A. COMMENTAIRE GENERAL SUR LA PUBLICATION

L'une des forces principales de notre étude dans le paysage des publications sur la normalisation et l'augmentation de données est que nous avons non seulement évalué nos méthodes sur la qualité du résultat en termes d'image, mais également en termes de généralisation sur un dataset biaisé. Autrement dit, contrairement à beaucoup d'études, nous avons monitoré l'objectif ultime de ce type de méthodes, à savoir le gain en termes de généralisation des résultats pour une tâche de classification.

Même si nos méthodes HEnorm et AugmentHE ne font pas disparaitre tous les biais qui existent entre les différentes cohortes, elles réussissent à contraindre le modèle à utiliser plus largement d'autres caractéristiques des images que la coloration afin de définir le grade histologique. La projection orthogonale en 2D de la représentation des patches dans l'espace latent du modèle ResNet avec ou sans HEnorm et AugmentHE est présentée dans la Figure 23. On observe une franche séparation des cohortes quand on n'utilise ni normalisation ni augmentation. L'effet de la normalisation seule consiste non seulement en un rapprochement des centroïdes de chaque cohorte, mais également en une agrégation des patches autour de ces centroïdes. L'effet de l'augmentation seule consiste également en un rapprochement des centroïdes de chaque cohorte, mais également dans une dispersion autour de ces centroïdes qui permet aux patches des différentes cohortes de se superposer. Enfin, l'effet combiné de la normalisation puis de l'augmentation est de rapprocher de manière plus marquée les centroïdes des différentes cohortes et d'appliquer une dispersion qui permet une superposition maximale des patches des différentes cohortes. En d'autres termes, la position d'un patch dans l'espace latent du modèle ResNet entraîné avec normalisation puis augmentation est moins corrélée à sa cohorte d'origine qu'avec les autres modèles.

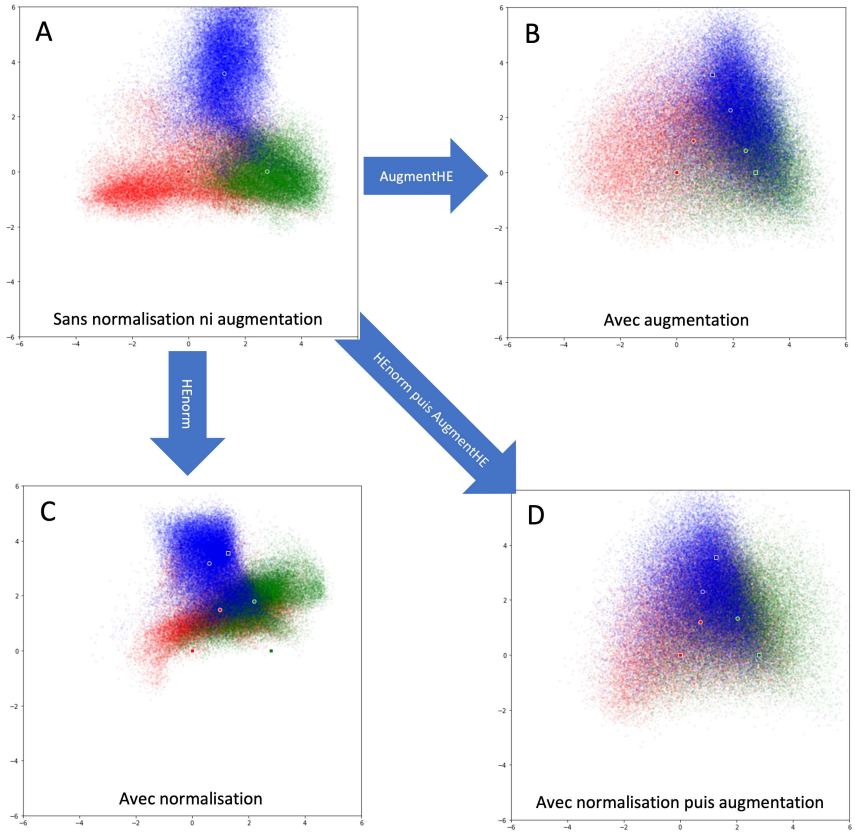


Figure 23 : Projection orthogonale en 2D de la représentation des patchs dans l'espace latent du modèle ResNet (A) entraîné sans normalisation ni augmentation, (B) avec AugmentHE seul, (C) avec HEnorm seul ; (D) avec HEnorm puis AugmentHE. Points rouges : PACS04, points verts : PACS05, points bleus : PACS08.

Avec l'objectif d'aborder les problèmes au plus près de leur source, nous pouvons également interroger la variabilité de coloration HE entre les laboratoires. Comme cela a été abordé dans l'introduction, l'origine naturelle de l'hématoxyline ainsi que la diversité des formulations et des méthodes de coloration expliquent en grande partie cette variabilité. En France, s'y ajoute une courante utilisation du safran ou de l'orange G qui colorent le collagène en jaune-orangé. Le point primordial concerne la qualité de cette coloration HE. Les structures d'assurance qualité en anatomie pathologique (telle que l'AFAQAP en France) proposent des tests d'évaluation externe de qualité qui permettent de s'assurer de sa bonne qualité technique et de la validité du protocole de coloration standard. Un pas vers une possible harmonisation des pratiques consisterait à mieux qualifier et « quantifier » la coloration HE des laboratoires. C'est l'objectif de plusieurs études qui considèrent centrale l'harmonisation de la coloration standard dans les laboratoires de pathologie à l'ère de la pathologie digitale [86,87]. De manière pragmatique, une telle évaluation dans les datasets utilisés pour l'entraînement de modèles d'IA pourrait permettre de produire des statistiques descriptives plus précises et de planifier des analyses de sous-groupes afin d'identifier d'éventuelles sources de biais.

B. AVANTAGES ET LIMITES DE L'OUTIL HENORM

Alors que le modèle d'augmentation de couleurs AugmentHE présente un intérêt évident et un effet majeur sur la généralisation des résultats, l'intérêt de l'outil HEnorm est plus marginal et peut être discuté. L'un des aspects intéressant d'HEnorm est qu'il peut être facilement répliqué et amélioré dans un autre laboratoire ou sur d'autres organes. En effet, il consiste à utiliser un dataset de lames d'un laboratoire de référence (dataset Normacolor dans la publication) afin d'entraîner un modèle à reproduire la coloration dudit laboratoire à partir d'une lame virtuelle d'entrée. Il faut noter que dans notre travail, le dataset Normacolor présentait une limite majeure : il était composé de seulement 700 lames de pathologie mammaire (bénigne ou maligne), ce qui est relativement restreint par rapport à ce qui pourrait être réalisé à l'échelle d'un laboratoire numérisant toute son activité de routine. On peut donc imaginer qu'un outil HEnorm entraîné à partir d'un dataset de plusieurs dizaines de milliers de lames serait plus efficace et aurait potentiellement un impact plus important sur la généralisation des résultats *in fine*, en association avec AugmentHE à l'entraînement, puis seul à l'étape d'inférence. Notons tout de même que l'excellent article de Tellez et collaborateurs arrivait aux mêmes conclusions que nous, à savoir que l'augmentation de données avait un impact beaucoup plus important sur la généralisation des modèles que la normalisation [88]. Finalement, comme toute méthode d'atténuation des biais, nos outils n'ont pas la prétention de faire disparaitre toute trace de biais lié à la coloration, mais de rendre les modèles médicalement plus pertinents, y compris dans des analyses de sous-groupes techniquement ou médicalement pertinents.

Comme cela a été abordé dans la discussion, les sources de biais sont multiples et peuvent intervenir tant dans la constitution des datasets, que dans l'entraînement des modèles et dans leur déploiement. Afin d'obtenir les meilleurs résultats, les méthodes d'atténuation des biais devraient être combinées et s'envisager à toutes les étapes du développement d'un modèle de machine learning [80].

C. AUTRES METHODES D'ATTENUATION DES BIAIS

L'atténuation des biais doit être appréhendée comme un processus multi-étapes, intervenant dès la phase de pré-traitement des données, puis lors de l'entraînement du modèle, puis sur le modèle entraîné. Dans une revue exhaustive du domaine, Max Hort et collaborateurs décrivent les nombreuses approches d'atténuation des biais existantes [89]. Dans le cadre de l'application au domaine de la pathologie digitale, nous retiendrons ici les méthodes qui semblent les plus pertinentes.

Pré-traitement des données (Pre-processing)

- **Perturbation** : entre autres les méthodes de normalisation (perturber les données de tous les groupes pour les rapprocher entre eux en modifiant des attributs de l'image qui n'ont pas de lien avec le label) et d'augmentation de données (accroître virtuellement le nombre d'exemples et la variabilité des données afin de représenter au mieux la variabilité observée dans la population cible).
- **Échantillonnage** : sur-échantillonnage ou sous-échantillonnage afin d'équilibrer les classes lors de la phase d'apprentissage et de validation.
- **Représentation** : ajout ou suppression de features. Inclut les notions de sélection de features, très fréquemment utilisées en machine learning conventionnel et plus rarement en deep learning [90]. Les approches de sélection de features sont extrêmement nombreuses et ne seront pas abordées ici.

Phase d'entraînement (In-processing)

- **Régularisation et pénalisation** : comprend les méthodes de bagging, de boosting et de régularisation. Ceci comprend notamment des hyperparamètres extrêmement importants en deep learning comme la *batch normalization*, le *dropout* et le *weight decay*.
- **Apprentissage adversarial** : consiste à entraîner plusieurs réseaux de neurones en parallèle. L'un réalise la tâche de classification tandis que l'autre identifie une source de biais. Une compétition entre les deux modèles s'opère afin que le classifieur atteigne de bonnes performances sans « profiter » du biais détecté par le réseau adverse. Dans notre exemple, il s'agirait d'associer un modèle de classification du grade histologique à un modèle de détection de la cohorte d'origine. Les performances obtenues par le réseau adverse agiraient négativement sur la fonction de coût du réseau de classification de grade.
- **Composition** : consiste à entraîner un modèle par catégorie de population. Ou dans notre cas, un modèle par cohorte.

Sur modèle entraîné (Post-processing)

- **Correction des images en inférence** : c'est ce que nous avons proposé en utilisant HEnorm lors de l'étape d'inférence. On rapproche ainsi les couleurs de la donnée d'entrée de celles que le modèle a rencontré pendant l'apprentissage.
- **Correction de la prédiction** : peut correspondre à un ajustement de seuil ou à définir des options de rejet (par exemple, un intervalle de scores de prédiction dans lequel le modèle ne se prononce pas). Ces méthodes de correction de la prédiction ne devraient

être envisagées qu'après vérification de la bonne calibration du modèle.

Enfin, on peut également interroger la propension de différents types de modèles d'IA au surapprentissage. En effet, des modèles proposant une meilleure représentation de la donnée pourrait être moins enclins au surapprentissage. A l'heure actuelle, on pourrait évoquer l'utilisation d'approches basées sur l'attention, telles que les *Vision Transformers*, ou des approches de *self-supervized learning*.

D. TRANSPARENCE ET ETHIQUE

D.1 MONITORER LES BIAIS

Beaucoup de dispositifs médicaux sont biaisés[13]. Il est important de le savoir afin d'affronter la question des biais avec éthique et intégrité scientifique. La particularité des biais associé à l'IA est qu'ils sont à l'origine d'erreurs difficiles à anticiper ou à prévenir. En imagerie médicale, la majorité des erreurs humaines provient soit d'un défaut de perception (par exemple, une lésion subtile, non vue par le médecin), soit d'une recherche incomplète d'anomalies dans une image entière, soit du syndrome de satisfaction qui consiste à diminuer sa vigilance après avoir identifié une première anomalie [91]. A l'inverse, les modèles d'IA ne sont pas soumis à des défauts de type recherche incomplète ou syndrome de satisfaction. Les modèles de deep learning, par nature, identifient des corrélations mathématiques complexes et opaques entre des données d'entrée et des prédictions. Cette puissance de calcul permet, dans le meilleur des cas, d'apprendre des patterns complexes et cliniquement pertinents, mais soulève le risque majeur d'identifier des corrélations fallacieuses, présentes dans le dataset d'apprentissage, par hasard ou du fait d'un biais systématique, mais ne correspondant pas au monde réel. D'autre part, on attend des modèles d'IA qu'ils soient performants en interpolation, c'est-à-dire dans le domaine représenté lors de la phase d'apprentissage, mais ils ne peuvent en aucun cas réaliser un processus d'extrapolation, c'est-à-dire interpréter des données en dehors du domaine représenté lors de la phase d'apprentissage.

Dans ce contexte, la description précise des données d'apprentissage, de validation et de test est un effort indispensable à fournir dans le sens de l'identification des biais. Les statistiques descriptives des datasets devraient non seulement intégrer des données en rapport avec la pathologie étudiée, mais également avec les variables clés de la phase pré-analytique et de l'acquisition des images. D'autre part, les études relatant les biais sociaux embarqués dans les modèles d'IA, y compris les études intersectionnelles, évoquent le besoin de définir la population

[13] https://www.theguardian.com/society/2021/nov/21/from-oximeters-to-ai-where-bias-in-medical-devices-may-lurk

ayant servi à la constitution des datasets sous des angles tels que l'origine ethnique ou le niveau de revenu [84]. En France, ces données sensibles ne sont en général pas collectées, et quand bien même elles le seraient, leur utilisation est très règlementée[14].

Au-delà de la description précise des données, les bonnes pratiques incluent la planification d'analyses de sous-groupes, tant sur des critères médicaux – *patient-specific subgroups* (par exemple : vérifier la validité d'un modèle par classes d'âge ou dans des sous-types tumoraux minoritaires), que techniques – *task-specific subgroups* (par exemple : vérifier la validité d'un modèle en fonction du scanner de lames utilisé, ou du type de coloration) [92]. Les sous-groupes basés sur des critères médicaux se forment classiquement à partir des données démographiques et cliniques présentées dans le tableau 1 des articles (d'où leur nom de *table 1 subgroup analysis*). Tandis que les sous-groupes techniques sont plus difficiles à constituer, nécessitent parfois des annotations supplémentaires, et leur nombre est théoriquement infini. C'est pourquoi il est conseillé de prioriser ces sous-groupes techniques sur la base des connaissances du domaine (dans le cas de la pathologie digitale, connaissance des étapes pré-analytiques et acquisition d'image par exemple) en fonction de leur risque et de leur criticité *via* une démarche AMDEC (Analyse des Modes de Défaillances, de leurs Effets et de leur Criticité). Idéalement, ces analyses de sous-groupes devraient être anticipées au même titre que les analyses de sous-groupes dans les essais thérapeutiques.

Une autre facette du monitoring des biais concerne la caractérisation des erreurs et leur impact clinique. L'analyse des erreurs du modèle est exploratoire par nature. Elle se base sur un examen systématique des faux positifs et des faux négatifs, ainsi qu'à des cas correctement prédits pour comparaison, afin d'identifier deux types d'erreurs : des erreurs « compréhensibles » (*predictive errors*) liées à une situation clinique particulière ou facteur confondant évident (et nécessitant potentiellement une analyse de sous-groupe dédiée), et des erreurs « surprenantes » (*surprise errors*) dans lequel l'examinateur ne comprend pas comment le modèle a pu échouer [92]. Cette analyse des erreurs devrait être couplée à l'utilisation d'outils d'explicabilité. Même si leur pertinence est débattue, et leur interprétation soumise au biais de confirmation, les cartes de salience (*saliency maps*) peuvent permettre de mieux appréhender l'origine des erreurs [93]. Comme cela sera discuté plus loin, cette notion cruciale d'interprétabilité des modèles devrait

[14] Selon les termes de l'article 6 de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés : "Il est interdit de traiter des données à caractère personnel qui révèlent la prétendue origine raciale ou l'origine ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale d'une personne physique ou de traiter des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.". Dans un rapport sur la mesure de la diversité et la protection des données personnelles daté de mai 2007, la CNIL se prononce contre la production de statistiques construites à partir d'une nomenclature de catégories "ethno-raciales" tout en présentant 10 recommandations (https://www.cnil.fr/fr/mesure-de-la-diversite-statistiques-ethniques-egalite-des-chances-les-10-recommandations-de-la-cnil) visant à améliorer la mesure de la diversité. Elle reconnaît par exemple la possibilité, dans le cadre de la statistique publique, d'études sur le ressenti des discriminations pouvant inclure des données sur l'apparence physique. De même, elle considère que, sous certaines conditions, l'analyse des prénoms, patronymes, nationalités et lieux de naissance des ascendants peut permettre de révéler des pratiques discriminatoires. [Source : https://www.vie-publique.fr/eclairage/19354-faut-il-elaborer-des-statistiques-ethniques]

autant que possible être envisagée *a priori* et non *a posteriori* [94].

L'impact clinique potentiel des erreurs des modèles d'IA est rarement exploré [95]. Il s'agit d'une démarche importante car toutes les erreurs ne se valent pas. Certaines n'auront aucune conséquence néfaste alors que d'autres peuvent être catastrophiques. Une telle évaluation nécessite une étroite collaboration entre data scientists et médecins, au service de la sécurité des patients et du déploiement des outils d'IA dans le domaine médical.

Enfin, les métriques permettant d'appréhender les biais sont nombreuses et devraient être adaptées à la question posée, aux objectifs cliniques du modèle et à des considérations éthiques [96]. Sans rentrer dans les détails, un modèle ne peut pas satisfaire toutes les conditions à la fois. Un compromis doit être trouvé entre un modèle équitable en termes, d'une part, de valeur prédictive positive ou négative ; d'autre part, en termes d'égalisation des chances (*equalized odds* – une mesure de l'équité des modèles d'apprentissage automatique).

Tout ou partie des points décrits ci-dessus permettant d'atténuer et de monitorer les biais des modèles d'IA sont partie intégrante du développement d'outils d'évaluation de la qualité des articles d'IA en médecine. Nous allons décrire quelques-uns de ces outils.

D.2 OUTILS D'EVALUATION DES ARTICLES D'IA

Ces outils constituent à la fois de bons guides pour l'analyse critique des articles scientifiques, mais également une checklist à suivre pour la réalisation de projets d'IA en médecine. Deux de ces outils sont détaillés ci-dessous : APPRAISE-AI et QUADAS-2.

D.2.1 APPRAISE-AI

APPRAISE-AI est un outil d'évaluation de la qualité méthodologique et de la robustesse des résultats de modèles d'IA pour l'aide à la décision clinique [97]. Il se présente sous forme d'une grille d'évaluation de 24 items regroupés en six axes permettant d'obtenir une note sur 100 : pertinence clinique (4 points), qualité des données (24 points), méthodologie (20 points), robustesse des résultats (20 points), qualité des informations communiquées (12 points) et reproductibilité (20 points). Parmi d'autres, on retrouve une évaluation du *data splitting*, des hyperparamètres du modèle, des biais et des erreurs et de leur impact clinique, ainsi que de la calibration du modèle. L'évaluation des biais comprend l'évaluation des performances du modèles dans différents sous-groupes relatifs aux patients ou aux tâches. Dans notre exemple, ceci impliquerait non seulement d'évaluer les performances du modèle par cohorte, mais également, par exemple, par statut ménopausique des patientes (groupes relatifs aux patientes) ou par sous-type histologique ou par immunophénotype (groupes relatifs aux tâches de

classification). Alors qu'elles constituent manifestement des éléments de bonnes pratiques, ces analyses de sous-groupes sont rarement réalisées (ou du moins communiquées) dans les articles d'IA en médecine. La majorité des articles produisant ce type d'analyses poussées sont des articles qui s'attachent spécifiquement à l'identification des biais [84,98].

D.2.2 QUADAS-2

QUADAS-2 est un outil d'évaluation de la qualité des études de performance diagnostique [99]. Il se base sur 4 items : la sélection des patients, le test index (outil diagnostique à tester), le test de référence, le flux et la temporalité entre les deux tests. Chaque item est abordé sous l'angle du biais potentiel et de l'applicabilité. Le risque de biais et les préoccupations en termes d'applicabilité sont notés « bas », « haut » ou « peu clair ». Dans une revue systématique de 12 articles comparant la validité des diagnostics en pathologie digitale vs au microscope, l'utilisation de l'outil QUADAS-2 a permis de révéler Deux études françaises récentes ont utilisé cet outil pour l'évaluation de la qualité des études évaluant des outils d'IA pour la prédiction 1) de l'instabilité microsatellitaire et des mutations de KRAS et BRAF dans des WSI de cancer colorectal [100] ; 2) du diagnostic de pathologies hépatiques tumorales, métaboliques ou inflammatoires [101]. Chacune de ces études identifie une proportion non négligeable d'études avec un haut risque de biais. Dans le premier article à propos des cancers colorectaux, sur 17 études incluses dans la revue, 11 avaient au moins un facteur de haut risque de biais et 6 n'en avaient aucun. Les risques de biais provenaient le plus souvent de de la sélection des patients (dataset de petite taille et/ou absence de cohorte de validation externe). Dans le second article à propos des maladies hépatiques, 42 études ont été évaluées. La majorité des études présentait au moins un item à haut risque de biais. Notons notamment qu'en pathologie non tumorale, la faible reproductibilité des critères diagnostiques ou pronostiques de référence était systématiquement associée à un haut risque de biais (biais d'annotation).
Une étude pan-organes s'est intéressée aux performances des tests diagnostiques basés sur l'IA [102]. Les domaines suivants étaient couverts : pathologies mammaire, cardiaque, dermatologique, hépatobiliaire, gastro-intestinale, urologique et autres. A l'aide de l'outil QUADAS-2, ils ont identifié que 99 % des études sur les 100 études analysées avaient au moins un item à risque de biais haut ou peu clair, ou soulevaient des préoccupations en termes d'applicabilité du modèle en pratique courante.

D'autres outils de ce type existent, tels que TRIPOD+AI et PROBAST+AI, ou encore QUADAS-C et QUAPAS, dérivé de QUADAS-2 mais dédiés aux études comparatives de performance diagnostique et aux études d' évaluation du pronostic, respectivement [103–105].

D.3 ETHIQUE

L'éthique va de pair avec le besoin de transparence. Il s'agit d'un très vaste sujet qui ne peut qu'être effleuré ici. Le premier biais associé au machine learning par l'équipe de Van Giffen est le biais social [83]. Il s'agit du biais inhérent aux biais sociétaux, notamment touchant des minorités ethniques ou des populations à faible niveau de revenus. Une reconduction, voire une possible amplification de biais sociétaux par les modèles d'IA a été démontrée par une étude intersectionnelle réalisée sur des modèles de screening de radiographies thoraciques (RT) [98]. Ces modèles de screening ont pour but de classifier avec une grande certitude les RT normales (optimisation de la valeur prédictive négative du test). Même si ces modèles atteignent de très bonnes performances, l'étude approfondie des erreurs du modèle (faux négatifs, correspondant cliniquement à un sous-diagnostic) permet d'identifier un biais important en défaveur de populations déjà souvent désavantagées dans la société. En l'occurrence, le taux de sous-diagnostics était plus élevé chez les femmes, les jeunes de moins de 20 ans, les patients noirs, les patients hispaniques et les patients « bénéficiant » de l'assurance Medicaid que dans les autres groupes (hommes, autres catégories d'âge, non noirs, non hispaniques, autres régimes d'assurance). De surcroit, ce taux de sous-diagnostic était additif : la combinaison de plusieurs facteurs de sous-diagnostic aboutissait à un taux encore plus important de sous-diagnostic.

On pourrait penser que les lames virtuelles sont exemptes de ce type de biais social. Cependant, des études montrent que la même vigilance est de mise dans le domaine de la pathologie digitale [84]. Les performances de modèles, y compris à l'état de l'art comme des modèles de fondation, semblent perpétuer des disparités entre des groupes démographiques (origine ethnique, niveau de revenu). Alors que l'origine ethnique des patientes en sénologie est directement associée à des épidémiologies distinctes, telle que par exemple la grande prédominance des cancers du sein triples négatifs en Afrique de l'Ouest, il semble potentiellement délétère de ne pas prendre en compte ce type de notion épidémiologique dans la définition de sous-groupes à l'étape d'évaluation des modèles [106,107].

Ces différents éléments vont également dans le sens d'une évaluation précise des limites des modèles en pathologie, afin de ne pas reproduire ou amplifier des déséquilibres déjà présents dans la société. Ceci passera par l'utilisation de modèles d'IA dans des essais cliniques prospectifs, qui permettront, au-delà des nombreuses preuves de concepts, d'atteindre un niveau satisfaisant de validation clinique des outils à disposition [23]. La transparence des résultats et la reproductibilité des études sont des éléments indispensables au concept de *Fairness in AI* [108–110].

IV. PERSPECTIVES

Le travail que nous avions prévu initialement, c'est-à-dire le développement d'un classifieur grade I vs grade III, puis l'évaluation de la valeur pronostique de ce classifieur sur des tumeurs de grade II RH+ HER2- a finalement été réalisé de bout en bout avec succès par l'équipe de Wang et collaborateurs [111].

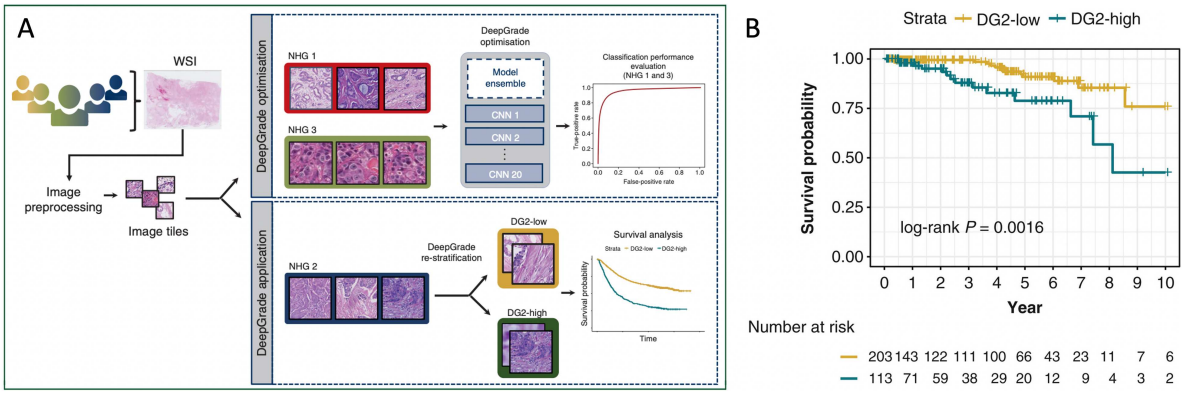


Figure 24 : Design (A) et résultats (B) de l'étude de Wang et al. [111].

Au lieu de nous décourager, ce fût l'occasion de nous interroger sur le véritable besoin médical sous-jacent dans ce domaine. Certes, une amélioration de l'estimation du pronostic des patientes avec de tumeurs de grade II RH+ HER2- est une avancée indiscutable. Mais que peut-on retirer réellement de cette information ? Et si ce modèle était lui aussi biaisé ? Un tel modèle accroit les connaissances des médecins sur les facteurs morphologiques influençant le pronostic ? Comment pourrait-on faire mieux ?

Les réponses à ces questions allaient bien sûr dans le sens d'un long chemin à parcourir vers plus d'interprétabilité, plus d'intégration des outils d'IA dans les essais cliniques prospectifs, plus de progrès en médecine au-delà de la simple optimisation de tâches [23].

En tant que pathologiste, l'interprétabilité des résultats des modèles me semble au centre des préoccupations. L'interprétabilité des modèles pourrait aussi bien participer de l'identification des biais, que de la confirmation ou de la découverte de corrélations et de liens de causalité, que de l'amélioration de la robustesse des modèles et de la confiance accordée aux outils d'IA en médecine [94,112]. Le défaut d'interprétabilité des modèles provient en grande partie du *gap sémantique* qui existe entre le fonctionnement des outils d'IA et le vocabulaire médical. Les perspectives amenées par ce projet vont dans le sens de la mise en place d'un langage commun entre l'homme et la machine. En pathologie, cela pourrait se faire concrètement par la segmentation préalable de multiples objets biologiques constituant les briques fondamentales

des tissus et permettant de construire des features de plus en plus complexes et adaptées au domaine dans une démarche de *feature engineering* [113]. Ces approches avaient quelque peu disparu depuis l'avènement du deep learning mais connaissent actuellement un renouveau très encourageant [114,115]. L'effort investi dans la création de features adaptées au domaine semble porter ses fruits. C'est dans cette philosophie que s'inscrit le projet APRIORICS que je porte depuis 2021.

A. PROJET APRIORICS

Le projet APRIORICS vise à développer un descripteur extensif et intelligible d'images microscopiques de cancer du sein. Au lieu de décrire un cancer du sein avec une quinzaine de paramètres comme le fait le pathologiste en pratique de routine, ce descripteur se baserait sur la segmentation de plusieurs objets biologiques de base (noyaux, cellules tumorales, mitoses, lymphocytes, macrophages, vaisseaux, in situ vs infiltrant) afin de produire une description plus complète de contenu de la lame virtuelle. Ce serait un préalable à l'entraînement de modèle d'IA basés sur des objets biologiques plutôt que sur des matrices de pixels.

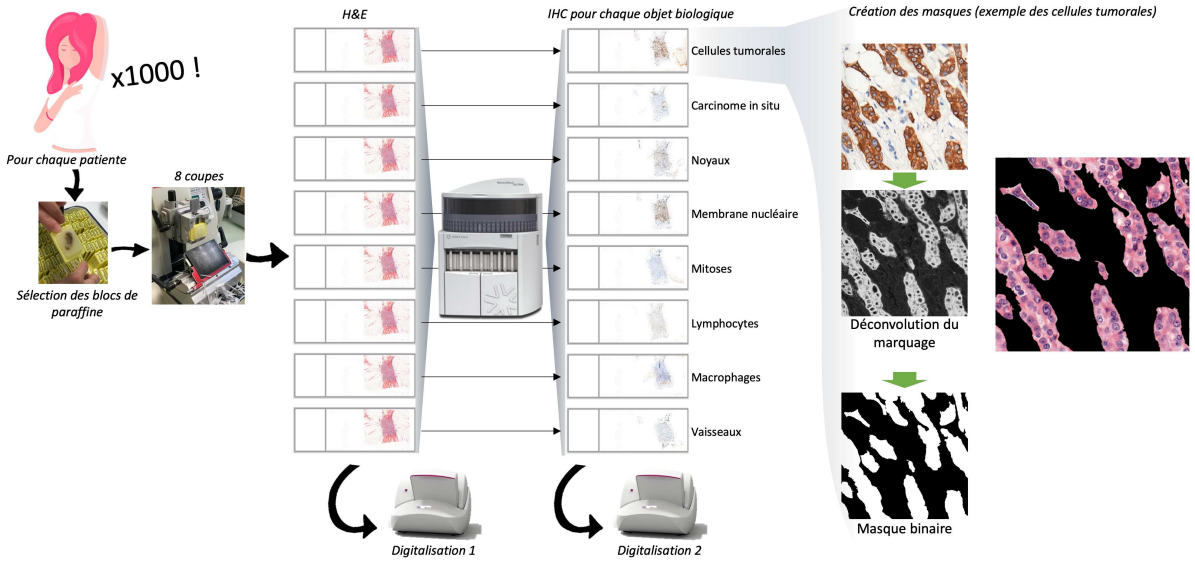


Figure 25 : Stratégie de production de données et d'annotation automatique des images par immunohistochimie dans APRIORICS.

La particularité de ce projet est d'avoir mis au point une méthode d'annotation automatique des images grâce à une utilisation détournée de l'immunohistochimie. Ainsi, neuf datasets en parallèle, de plus de 1000 lames chacun, sont en cours d'annotation par cette méthode (Figure

25). Notre méthode basée sur l'immunohistochimie, et donc sur l'expression de protéines spécifiques dans les tissus, permet de s'affranchir du biais d'annotation que tout pathologiste pourrait introduire en annotant manuellement des lames. D'autre part, cette méthode permet d'annoter une quantité très importante d'objets à moindre coût (Figure 26).

Dans une optique d'ouverture et dans le cadre du soutien de ce projet par le Health Data Hub, l'un des objectifs est de rendre cette cohorte annotée accessible à la communauté via le catalogue du Health Data Hub.

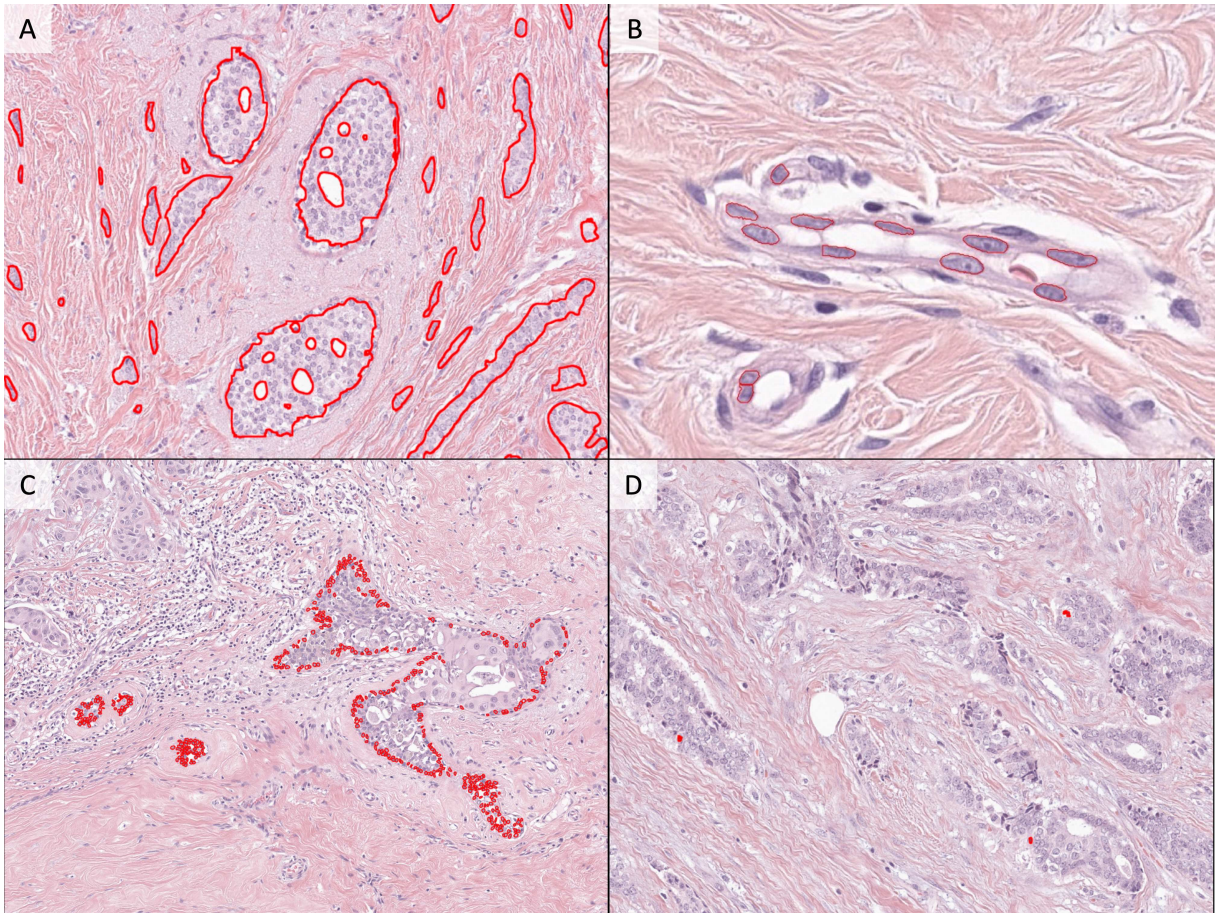


Figure 26 : Exemple d'annotations automatiques obtenues dans le cadre du projet APRIORICS. (A) Annotation des cellules épithéliales (grossissement x120) ; (B) Annotation des noyaux des cellules endothéliales (grossissement x400) ; (C) Annotation des noyaux des cellules myoépithéliales (grossissement x120) ; (D) Annotation des mitoses (grossissement x 120). Les traits rouges représentent les structures annotées.

B. PARTICIPATION AU PROJET MALO

Dans le cadre d'une extension inattendue du projet APRIORICS, nous avons pu interagir longuement avec un philosophe de la technique (Pr Xavier Guchet) et une épistémologiste (Océane Fiant). Le projet MaLO avait pour but de suivre un certain nombre de projet d'IA en

médecine (dont APRIORICS) afin de réfléchir sur les impacts potentiels de ces nouvelles technologies sur les professions médicales.

Les longs entretiens associant Xavier Guchet, Océane Fiant, Robin Schwob (ingénieur du projet APRIORICS) et moi-même ont été très fructueux. Les sujets abordés étaient multiples et peu à peu, la question de l'interprétabilité des modèles s'est imposée et a fait l'objet d'une publication dans la revue Réseaux (article accepté pour publication). Ce fût une expérience très enrichissante et les interactions avec Océane Fiant se poursuivent et pourraient aboutir à d'autres travaux dans le futur, à la frontière de la médecine, de l'épistémologie et de la philosophie.

AU-DELÀ DE L'EXPLICABILITÉ.
Étude de la conception d'une intelligence artificielle intelligible en anatomie et cytologie pathologiques

Océane FIANT (oceane.fiant@univ-cotedazur.fr), *GREDEG, Université Côte d'Azur et Costech, Université de technologie de Compiègne*

Camille FRANCHET (franchet.camille@iuct-oncopole.fr), *Département de pathologie, Institut universitaire du cancer Toulouse-Oncopole*

Robin SCHWOB (schwob.ro@chu-toulouse.fr), *Centre hospitalier universitaire de Toulouse*

***Résumé* :**

L'explicabilité des intelligences artificielles (IA) est souvent présentée comme un élément indispensable à l'appropriation de ces technologies par les médecins. Toutefois, son approche habituelle vient se heurter à deux écueils : d'abord, l'absence d'ancrage dans des situations professionnelles réelles, entraînant un décalage entre les solutions proposées et les attentes concrètes des utilisateurs ; ensuite, une focalisation excessive sur le fonctionnement des IA, au détriment d'autres aspects de leur conception pouvant également induire un défaut d'intelligibilité des résultats. Cet article se concentre sur un projet d'IA en anatomie et cytologie pathologiques qui propose une perspective inédite sur l'intelligibilité des IA, mettant l'accent sur la construction de leur « vérité de terrain ». L'examen de la stratégie mise en œuvre par ce projet plaide alors en faveur d'une approche contextuelle de l'appropriabilité des IA, permettant de développer des solutions réellement alignées sur les attentes des professionnels de santé.

***Mots-Clés* :** explicabilité ; appropriation ; intelligence artificielle ; anatomie et cytologie pathologiques ; ensemble de données ; vérité de terrain

Dans la littérature, l' « explicabilité » des intelligences artificielles (IA), entendue comme la capacité à élucider les mécanismes par lesquels celles-ci formulent leurs résultats, est souvent présentée comme un élément essentiel à l'intégration effective de ces technologies dans les pratiques médicales (Amann *et al*., 2020 ; Gerdes, 2024). Ainsi, le Comité consultatif national d'éthique pour les sciences de la vie et de la santé et le Comité national pilote d'éthique du numérique, dans leur avis conjoint sur les enjeux éthiques du diagnostic médical et de l'IA, soulignent la nécessité de « créer les conditions de la confiance en incitant les développeurs à fournir un certain niveau d'explicabilité » de leurs systèmes (2022). En effet, l'opacité des IA, qui se traduit par une difficulté à identifier quelles caractéristiques des données orientent leurs prédictions et pour quelles raisons, est perçue comme un obstacle à leur acceptation par les professionnels de santé[1]. Selon la *Defense Advanced Research Projects Agency*, pionnière dans le domaine de l'*explainable artificial intelligence* (*XAI*), produire des modèles plus explicables devrait ainsi permettre aux utilisateurs de « développer une compréhension, une confiance raisonnable et des capacités de pilotage effectives de la nouvelle génération d'outils d'IA[2]» (Gunning, 2017).

L'explicabilité des IA est de ce fait considérée comme un défi majeur à relever. En effet, elle doit garantir la sécurité des patients sujets d'une prédiction algorithmique en mettant les médecins en capacité de rendre compte des résultats produits par les IA et d'en évaluer la pertinence clinique (Doshi-Velez *et al*., 2019 ; Barredo Arrieta *et al*., 2020 ; Lebovitz *et al*., 2022). L'explicabilité pourrait également avoir un intérêt heuristique, en favorisant la découverte de corrélations inconnues jusqu'alors, ouvrant ainsi de nouvelles pistes d'investigation scientifique (Herzog, 2022).

Pour rendre les IA explicables, deux stratégies sont le plus souvent proposées : il s'agit, d'une part, de l'utilisation de modèles dits « interprétables par nature », qui sont généralement définis comme des modèles dont les opérations sont compréhensibles par un être humain (Lipton, 2017b). D'autre part, l'*XAI* s'appuie sur l'utilisation de techniques dites « *post-hoc* », qui

[1] Une caractéristique des données (*feature*) est une propriété d'entrée utilisée par un modèle d'apprentissage automatique pour prédire. Par exemple, en analyse d'images, les caractéristiques présentes dans les données sont des pixels ou des ensembles de pixels.

[2] « to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners » (traduction auteurs).

fournissent des informations simplifiées sur le fonctionnement de modèles opaques[3].

Cependant, un examen de la littérature révèle deux principales lacunes. Premièrement, il apparaît que l'explicabilité est souvent abordée de manière abstraite, c'est-à-dire sans ancrage suffisant dans des contextes d'application concrets (Lipton, 2017a). Cette tendance peut conduire les développeurs à concevoir des méthodes explicatives qui reflètent davantage leur propre compréhension de ce qui constitue une explication pertinente, plutôt que celle des utilisateurs finaux (Miller *et al*., 2017). Cela peut aboutir à des solutions qui ne répondent pas adéquatement aux attentes de ces derniers (Ghassemi *et al*., 2021)[4].

Deuxièmement, affirmer que l'appropriabilité des IA par les médecins a pour condition *sine qua non* la possibilité, pour eux, de comprendre les mécanismes par lesquels ces systèmes produisent leurs résultats (ce qui est proprement la définition de l'explicabilité) est réducteur. Cette focalisation sur le fonctionnement interne des modèles tend à occulter le fait que des aspects de ces IA autres que leurs mécanismes internes sont de nature à conditionner leur appropriation par les praticiens. Par exemple, les choix techniques qui sous-tendent la construction des ensembles de données, appelés « vérités de terrain », utilisés pour l'entraînement et la validation des modèles, ainsi que le caractère plausible des résultats produits par les logiciels d'IA, conditionnent pour une grande part l'intelligibilité de ces systèmes pour les médecins, et donc leur appropriabilité dans les pratiques médicales (Jaton, 2017 ; Ratti et Graves, 2022 ; Gaglio et Loute, 2023).

---

[3] Les techniques *post-hoc* peuvent comprendre la création d'un second modèle, simple et interprétable, qui reproduit le comportement du modèle original en se fondant sur ses entrées et ses sorties. Elles peuvent également analyser l'importance de chaque caractéristique d'entrée pour la prédiction finale. Ces différentes techniques peuvent non seulement viser à expliquer des prédictions spécifiques (elles sont alors dites « locales »), mais également le fonctionnement du modèle (elles sont alors dites « globales »).

[4] Marzyeh Ghassemi *et al*. (2021) illustrent cette problématique en montrant l'échec des cartes de saillance, une méthode d'explication populaire en médecine, à éclairer effectivement la décision clinique. Les cartes de saillance permettent de visualiser les zones d'une image prises en compte par un modèle d'analyse d'images pour formuler son résultat. Toutefois, ces cartes ne précisent pas sur quelle(s) caractéristique(s) des données présente(s) dans ces zones le modèle a fondé sa prédiction. En conséquence, les cartes de saillance offrent au praticien une information partielle, l'obligeant à interpréter lui-même la pertinence médicale du résultat. Ghassemi *et al*. (2021) mettent en évidence le risque d'erreur de décision inhérent à cette méthode.

Ainsi, des enquêtes qualitatives menées auprès de médecins montrent que l'explicabilité n'est pas unanimement perçue comme une condition essentielle à l'intégration des IA dans les pratiques médicales (Drogt *et al.*, 2022 ; Gaglio et Loute, 2023).

Dans le prolongement de ces études, cet article soutient que l'explicabilité, telle que définie par l'*XAI*, n'est pas indispensable pour que les praticiens comprennent les résultats des IA, et donc pour l'intégration effective de ces technologies en médecine. Cependant, il va plus loin en soulevant une question épistémologique : sous quelles conditions les résultats des intelligences artificielles peuvent-ils être intelligibles pour les praticiens, s'il ne s'agit plus de rendre les modèles explicables ? La thèse défendue ici est que les résultats des modèles ne sont intelligibles que s'ils intègrent les connaissances biologiques et médicales des praticiens. Cela pose un problème particulier dans le cas des modèles de *deep learning*, qui, précisément, ne s'appuient sur aucune connaissance préalable (qu'il s'agisse de règles de décision, ou autres).

À partir de l'étude d'un projet d'IA en anatomie et cytologie pathologiques (ACP), une spécialité médicale centrale dans la prise en charge des cancers, cet article propose de montrer que la question de l'intelligibilité des IA, qui est l'une des conditions de possibilité de l'appropriation de ces dispositifs par les médecins, se rapporte à la question de savoir comment amener des connaissances (en l'occurrence, biologiques et médicales) dans les modèles. Il s'agit de montrer, à travers ce cas précis, que l'explicabilité des mécanismes internes des modèles utilisés n'est pas une question centrale pour le médecin qui développe ces outils.

Dans cette optique, la première partie de l'article s'attachera à montrer que le problème auquel le projet étudié cherche à répondre, ainsi que les éléments qui s'y rapportent, impliquent une compréhension de l'intelligibilité des IA qui dépasse le simple fait d'élucider le fonctionnement de celles-ci. La seconde partie présentera l'approche adoptée dans le cadre de ce même projet pour produire des IA dont les résultats font sens pour le praticien. Cette étude de cas permettra de montrer que la question de l'intelligibilité, et plus largement celle de l'appropriabilité des IA, n'a de pertinence que si elle est posée dans des contextes d'utilisation définis, caractérisés par des problématiques d'usage spécifiques. De ce fait, cet article s'inscrit dans la continuité des travaux qui, ces dernières années, ont montré que les défis associés à l'IA ne peuvent être pleinement compris qu'à partir d'une étude des environnements spécifiques où ces technologies sont mises en œuvre, ce qui

implique une étude approfondie de ces environnements (Mateescu et Elish, 2019 ; Sendak *et al*., 2020).

Ancrage théorique, terrain et méthode

Sur le plan théorique, cet article s'inscrit dans le « tournant empirique » en philosophie des techniques. Refusant de considérer la technique d'un point de vue purement spéculatif, les philosophes relevant de ce courant élaborent leurs questionnements philosophiques à partir de cas d'innovations concrets (en cours ou passés). Très inspirés par les *science and technology studies*, les auteurs des références qui orientent la démarche d'ensemble de l'article, notamment Peter-Paul Verbeek, formulent leurs analyses philosophiques sur la technique à partir de la technique (*from technology*), c'est-à-dire en examinant la façon dont les dispositifs sont conçus et déployés ; ils ont aussi l'ambition de proposer des analyses pour la technique (*for technology*), selon une démarche de recherche-intervention que Verbeek appelle « *technology accompaniment* » : le philosophe ne se contente pas d'observer, il veut aussi aider à orienter les processus d'innovation (2010 ; 2022).

Le projet étudié dans cet article, nommé « Apprentissage Profond Renforcé par l'Immunohistochimie pour la Requalification d'Images du Cancer du Sein » (APRIORICS), a bénéficié du soutien financier de Bpifrance et de l'appui technique du *Health Data Hub*. Lancé en 2021 et se poursuivant en 2024, ce projet est mené à l'Institut Universitaire du Cancer Toulouse-Oncopole (IUCT-O) par le Dr Camille Franchet, pathologiste spécialisé en sénologie, et Robin Schwob, *data-scientist*.

Ce projet vise à constituer un ensemble d'images annotées de tumeurs cancéreuses du sein. Ces données serviront à entraîner des réseaux de neurones convolutifs à identifier automatiquement les différents composants de la tumeur sur des lames d'histologie numérisées, tout en garantissant l'intelligibilité des résultats.

L'étude de ce cas a été réalisée par Océane Fiant, dans le cadre d'un projet de recherche financé par l'Institut national du cancer de 2021 à 2024. Elle s'est appuyée sur une enquête qualitative impliquant trois entretiens semi-directifs d'une durée de deux heures chacun, menés avec le Dr Camille Franchet et Robin Schwob, de septembre 2022 à mars 2023. Le premier entretien (septembre 2022) a été réalisé avec le Dr C. Franchet. Il avait pour but

d'appréhender les objectifs et les contours d'APRIORICS. Les deux entretiens suivants (octobre 2022 et mars 2023) ont été réalisés avec le Dr C. Franchet et R. Schwob. Ils ont porté sur des questions spécifiques telles que le développement, l'optimisation et la validation des modèles d'IA, l'annotation des données, les biais, l'effet attendu sur les pratiques d'analyse du pathologiste, ainsi que sur l'intelligibilité des résultats. Ces discussions ont été complétées par des démonstrations des modèles sur des exemples issus de l'ensemble d'entraînement. Elles ont été enrichies par de nombreux échanges informels, incluant des réunions et des correspondances par e-mails, jusqu'en juin 2024.

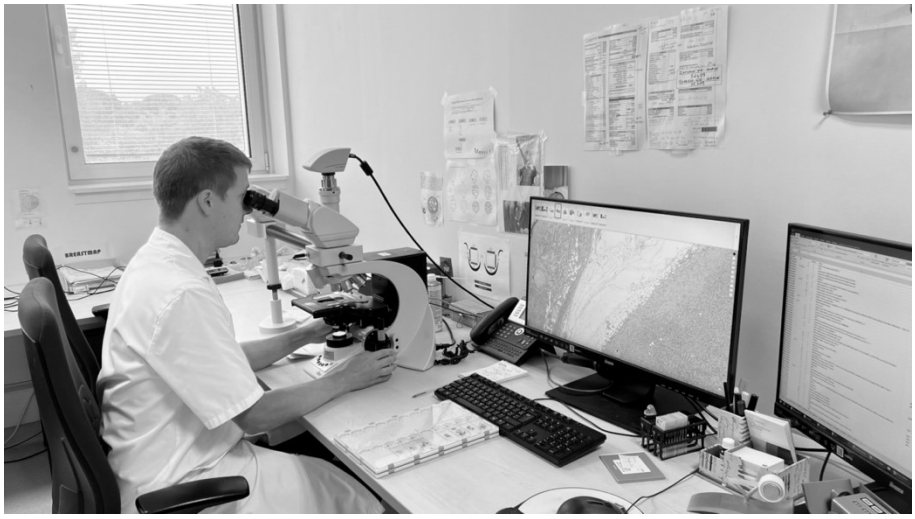
AUGMENTER LA CAPACITÉ D'ANALYSE DU PATHOLOGISTE GRÂCE À DES IA BIOLOGIQUEMENT PERTINENTES

ACP et IA : de l'apprentissage automatique traditionnel à l'apprentissage profond

L'ACP est une spécialité médicale qui se consacre à l'étude des anomalies provoquées par ou associées à des maladies sur les organes, les tissus et les cellules. Elle intervient notamment dans la prise en charge des cancers, où elle fournit aux cliniciens des informations capitales pour le diagnostic, le pronostic et la prise en charge thérapeutique. L'examen anatomopathologique porte sur des étalements de cellules isolées (cytopathologie) ou sur des prélèvements tissulaires (histopathologie) et associe étude morphologique, immunohistochimique et moléculaire de ces prélèvements[5].

[5] La morphologie est l'étude des structures cellulaires et tissulaires présentes dans un échantillon. L'immunohistochimie, quant à elle, est une technique permettant d'y localiser des protéines caractéristiques de certains événements cellulaires et d'en examiner l'expression. Enfin, des techniques issues de la biologie moléculaire, telles que le séquençage nouvelle génération et l'hybridation *in situ*, permettent la caractérisation moléculaire de cet échantillon.

Figure 1. Pathologiste combinant dans sa pratique examen microscopique traditionnel et analyse de lames virtuelles



Source : Dr Camille Franchet.

Depuis peu, dans les laboratoires qui ont mis en œuvre la « pathologie digitale », l'examen histopathologique peut être réalisé à partir de lames numérisées, appelées « lames virtuelles » (ou *whole slide images*). La pathologie digitale est apparue dans les années 2000 (Pantanowitz *et al*., 2018). Elle renvoie à l'acquisition, la gestion, le partage et l'interprétation de lames virtuelles – principalement histologiques – grâce à des microscopes motorisés ou des scanners de lames (Hanna et Pantanowitz, 2017). Pendant longtemps, son usage en clinique est resté très limité, et ce, en raison des coûts directs et indirects associés à la mise en place d'un flux de travail dédié dans les laboratoires. Néanmoins, aujourd'hui, dans la mesure où elle est une condition *sine qua non* pour le développement et l'usage de solutions à base d'IA en ACP, la conversion des laboratoires au numérique tend à s'accélérer (Drogt *et al*., 2022)[6].

[6] En témoigne le lancement en juin 2023 d'une mission sur la pathologie digitale, dont les objectifs sont notamment de faciliter la transition numérique de la spécialité et d'y amorcer le déploiement d'outils d'IA (Ministère du Travail, de la Santé et des Solidarités, 2023). Il convient de souligner que la pathologie digitale offre aux pathologistes la possibilité de travailler à distance et de partager des images avec les cliniciens, ce qui constitue également des arguments forts en sa faveur.

Dès son apparition, la numérisation de l'examen histopathologique a suscité un intérêt pour l'application des techniques issues du champ de la vision par ordinateur (*computer vision*) à l'ACP[7]. Initialement, les solutions conçues pour assister le pathologiste dans l'interprétation des lames virtuelles reposaient sur l'usage de modèles d'apprentissage automatique qualifiés de « traditionnels » par la communauté informatique, par contraste avec les modèles à base d'apprentissage profond apparus dans les années 2010 (Cardon *et al*., 2018). Ces modèles traditionnels nécessitent, pour accomplir une tâche donnée, que le *data scientist* identifie et extraie manuellement des données d'entrée les caractéristiques pertinentes pour réaliser la tâche en question. Or, cette approche rencontre rapidement ses limites lorsqu'il s'agit de l'appliquer à des lames virtuelles. Il s'agit en effet d'images très complexes, contenant des centaines de milliers, voire des millions d'objets biologiques. En outre, les objets d'une même classe sont hétérogènes. Par exemple, l'apparence d'une cellule en cours de mitose (taille, forme, texture, ton) sur une lame peut varier considérablement en fonction de la phase de division cellulaire dans laquelle celle-ci se trouve, ainsi qu'en fonction des méthodes employées pour préparer la lame (fixation, coloration, etc.)[8]. De fait, analyser ces images à l'aide de modèles d'apprentissage automatique traditionnels implique un fastidieux travail de la part du *data scientist* pour modéliser cette complexité, travail qui nécessite au demeurant la participation active des pathologistes, en tant qu'experts du domaine (van der Laak *et al*., 2021).

Malgré le développement de librairies informatiques facilitant l'extraction des caractéristiques, telles que *scikit-image*, l'approche traditionnelle a été supplantée, à partir des années 2010, par celle à base de réseaux de neurones convolutifs – qui constituent une catégorie de réseaux de neurones profonds. En effet, ceux-ci automatisent l'extraction des caractéristiques pertinentes des images, tout en démontrant des performances supérieures à celles de l'approche traditionnelle. Leur architecture se compose de deux parties : d'abord, une partie convolutive qui applique une série de filtres à l'image pour y détecter des formes et des textures, puis une partie classificatoire, qui

[7] La vision par ordinateur est un domaine de la science informatique qui porte sur l'analyse d'images naturelles.

[8] La mitose est le processus de division cellulaire au cours duquel le noyau d'une cellule somatique se divise pour former deux nouveaux noyaux, chacun contenant un ensemble complet et génétiquement identique de chromosomes diploïdes (c'est-à-dire, deux jeux de chromosomes). Ce processus se déroule en plusieurs étapes : prophase, prométaphase, métaphase, anaphase et télophase, au cours desquelles l'apparence de la cellule subit des changements significatifs.

combine ces caractéristiques pour classer l'image selon des catégories préétablies.

L'IA pour enrichir la capacité d'analyse du pathologiste

Dans le cadre du projet APRIORICS, les réseaux de neurones convolutifs sont employés pour faire de la segmentation d'images. Cette technique, qui existe depuis les années 1970, permet d'isoler et de localiser des objets dans une image[9]. Appliquée à des lames virtuelles, elle facilite notamment certaines tâches de quantification (mesure de la taille des cellules, de leur nombre, etc.)[10]. En cela, elle constitue à l'échelle internationale un axe de recherche important dans le domaine de l'analyse d'images histologiques (Wang, 2019).

L'objectif qui sous-tend l'usage de ces modèles dans le projet est de mettre le pathologiste en capacité d'exploiter plus amplement l'information présente dans son principal matériau d'étude, les lames histologiques. En effet, dans le cancer du sein notamment, l'information extraite de celles-ci lors de l'examen histologique ne porte que sur quinze critères, permettant de caractériser la tumeur en un temps limité (un quart d'heure pour les cas complexes) et de déterminer la prise en charge thérapeutique la plus adaptée. L'information contenue dans les lames est évidemment plus riche que ce que laissent apparaître ces quinze critères qui, au demeurant, concernent essentiellement les cellules tumorales :

> « Dans une lame de verre standard représentative d'un cancer du sein, on a souvent plus d'un million de cellules sous les yeux et une tumeur, c'est complexe. Ce n'est pas seulement un paquet de cellules tumorales, on a […] une interaction entre des cellules tumorales et l'environnement, l'immunité, la structure, les vaisseaux, etc. » (Pathologiste, entretien septembre 2022)

Par exemple, pour déterminer le grade d'une tumeur, c'est-à-dire son agressivité, les pathologistes se fondent notamment sur un critère nommé « index mitotique ». Ce dernier représente le nombre de cellules en mitose par mm^2 dans l'échantillon tissulaire. Le nombre de mitoses observées dans un

[9] Pour ce faire, les pixels possédant des propriétés similaires (couleur, intensité) se voient associer une même étiquette. L'image d'entrée est alors transformée en une carte de segmentation délimitant les contours de l'objet d'intérêt.

[10] Comme le soulignent Chiara Carboni *et al*. (2023), la pathologie digitale encourage des pratiques d'analyse quantitatives, une tendance que l'essor de l'IA est susceptible de renforcer en automatisant les tâches de quantification.

échantillon tumoral est un indicateur clé de la prolifération de la tumeur : un nombre élevé de mitoses suggère une croissance rapide de la tumeur, impliquant un pronostic moins favorable. Dans la mesure où le comptage exhaustif des mitoses sur toute la surface de la lame est irréalisable, l'index mitotique est estimé à partir d'une zone représentative d'environ 3 mm^2, examinée au grossissement maximal du microscope (x400). Cette zone est choisie pour sa forte activité mitotique, censée refléter le « comportement » global de la tumeur. Toutefois, si l'IA était capable d'inventorier l'intégralité des mitoses présentes sur la lame, elle permettrait d'obtenir une image plus précise de ce « comportement »[11].

L'usage de l'IA pour décrire de manière exhaustive le tissu tumoral pourrait donc, d'une part, améliorer l'exploitation des critères diagnostiques et pronostiques existants[12]. D'autre part, selon le pathologiste impliqué dans le projet, elle pourrait également faciliter l'identification de nouveaux critères, qui n'étaient pas exploitables auparavant en raison des limites des méthodes traditionnelles :

> « Pour ce qui est des relations entre les cellules, s'il y a plus de trois critères intriqués, on est coincés, alors que la machine... Par exemple […] quand on regarde l'infiltration de la tumeur par les cellules immunitaires, on peut regarder où se situent les cellules immunitaires : entre les amas de cellules cancéreuses ou à l'intérieur, vraiment au contact des cellules cancéreuses[13]. Et si le bon critère, celui qui est intéressant, c'est une histoire de proportion de cellules qui sont au contact par rapport à celles qui ne sont pas au contact, à l'œil, on ne sait pas l'évaluer. » (Pathologiste, entretien octobre 2022)

En somme, pour reprendre la distinction de François Sigaut, l'IA apparaît ici à la fois comme un « outil auxiliaire », améliorant l'analyse du pathologiste, et comme un « outil nécessaire », lui offrant la possibilité d'explorer des aspects qui lui étaient auparavant inaccessibles (2012).

---

[11] Non seulement le calcul de l'index mitotique se fonde sur une extrapolation, mais il est également variable entre les pathologistes. Des études ont en effet montré que, pour une même lame, la sélection de la zone utilisée pour le comptage varie entre pathologistes, tout comme le nombre de mitoses comptées dans la même zone (van Diest *et al*., 2004).

[12] Ainsi, récemment, la segmentation d'images a permis d'affiner le grade histologique des cancers du sein, un critère utilisé par les pathologistes pour évaluer, entre autres, l'agressivité de la tumeur (Amgad *et al*., 2024).

[13] L'évaluation de l'infiltration tumorale par les cellules immunitaires est utilisée à des fins pronostiques, ainsi que pour prédire la réponse à certains traitements.

Sous ces deux dimensions, l'IA contribuerait à enrichir significativement la capacité d'analyse du pathologiste, améliorant ce faisant sa précision diagnostique :

> « quand la machine nous chiffre tout, on a des surprises, des choses auxquelles on ne s'attendait pas et qui peuvent avoir un poids dans une décision, dans une sensibilité à des traitements » (Pathologiste, entretien octobre 2022)

Par-là même, l'IA pourrait permettre à ce professionnel de santé de recouvrer un rôle central dans des décisions médicales qui lui ont échappé au profit d'approches concurrentes, issues de la génomique.

Il convient en effet de noter que l'ACP a récemment vu ses techniques et, par conséquent, son rôle dans la prise en charge des cancers, remis en question par l'essor d'une médecine de précision fondée sur l'analyse du génome tumoral. Bien sûr, les approches issues de la génomique n'ont pas remplacé l'expertise du pathologiste de façon pure et simple (Bergeron *et al*., 2021). Néanmoins, force est de constater que les connaissances qui en émanent orientent désormais certaines décisions (Bourret *et al*., 2011)[14]. Ainsi, bien loin de s'inscrire inévitablement dans la problématique du remplacement de l'expertise du médecin, l'IA peut être enrôlée par certaines professions médicales pour renforcer leur expertise face à des expertises concurrentes.

La difficulté de conjuguer apprentissage profond et prédictions biologiquement signifiantes

Néanmoins, le pathologiste s'inquiète du caractère opaque des réseaux de neurones convolutifs. En effet, la manière même dont ces derniers extraient de l'information des lames virtuelles risque de faire perdre tout sens biologique à cette information :

[14] Cela est notamment le cas de la décision thérapeutique dans certains des cancers du sein hormonodépendants de stade précoce qui ne surexpriment pas le récepteur 2 du facteur de croissance épidermique. Généralement de bon pronostic, ces cancers sont majoritairement traités par hormonothérapie néoadjuvante et chirurgie. Cependant, un faible pourcentage présente un risque de récidive post-traitement, qui justifie la mise en place d'une chimiothérapie adjuvante. Ce risque est apprécié en fonction de critères cliniques et anatomopathologiques (âge, grade de la tumeur, etc.), mais il arrive que ces critères soient inopérants. Dans ces situations, sur la base des recommandations de bonne pratique, le cancérologue présent à la réunion de concertation pluridisciplinaire demandera la réalisation d'une signature moléculaire (en général, Oncotype DX©). Celle-ci indiquera le risque de la patiente et tranchera par conséquent en faveur ou non de la chimiothérapie adjuvante.

> « L'information extraite doit être intelligible pour un biologiste, pour un pathologiste, pour un oncologue. Extraire de l'information comme le font les réseaux de neurones convolutifs, c'est-à-dire en enchaînant de nombreuses convolutions afin d'extraire des caractéristiques distinctives, mais totalement incompréhensibles, des images, cela nous éloigne beaucoup de la compréhension des mécanismes biologiques sous-jacents. » (Pathologiste, entretien septembre 2022)

Comme le montre cette citation, aux yeux du pathologiste, le problème de l'« opacité » des réseaux de neurones convolutifs tient avant tout à l'absence de relation entre les résultats et un corpus établi de connaissances biologiques et médicales. En effet, comme indiqué précédemment, les réseaux de neurones convolutifs requièrent peu de prétraitement pour fonctionner. De fait, la manière dont ils analysent une image ne fait intervenir aucune connaissance préalable de la biologie et de l'histologie. Ce sont plutôt les transformations successives qu'ils font subir à l'image qui leur permettent de tirer une information de celle-ci. Or, la logique qui sous-tend ces transformations est difficilement accessible. Par conséquent, en situation d'usage, le pathologiste se trouverait dans l'incapacité de savoir si les objets mis en évidence par le modèle sont réellement des objets biologiques, ou s'ils résultent par exemple de la présence d'artefacts dans ses données d'entraînement[15]. Ainsi, le fait que le fonctionnement des modèles soit opaque n'est pas en soi problématique ; le problème est que cette opacité se paye du prix d'une décorrélation entre les résultats et les connaissances du médecin.

Cette problématique est rendue d'autant plus évidente par une expérimentation menée par l'équipe, et qui consistait à comparer, sur les mêmes cas cliniques, les résultats d'un réseau de neurones convolutif et ceux d'un modèle dit « interprétable » (XGBoost), c'est-à-dire explicable au sens de l'*XAI*. Ce type de modèle permet de calculer l'importance de chaque caractéristique d'entrée pour la prédiction finale. Au cours de l'expérience, le réseau de neurones convolutif traitait une image de la tumeur comme entrée, tandis que l'algorithme interprétable disposait d'un tableau reprenant les variables extraites de l'image, telles que le nombre de mitoses par mm^2.

[15] Dans une image histologique, un artefact est une anomalie résultant non pas d'une pathologie sous-jacente, mais d'altérations survenues lors du prélèvement, de la préparation ou de la numérisation de l'échantillon. Par exemple, il peut s'agir de bulles d'air emprisonnées pendant le montage de l'échantillon sur la lame.

> On voit bien qu'il y a des choses où, si on essaie de comprendre ce que fait le réseau, on a des intuitions qui sont bonnes et des intuitions qui sont fausses. On va penser : « ah oui, je pense que mon modèle a repéré telle chose dont on sait biologiquement que c'est associé à l'agressivité de la tumeur », par exemple. Et puis quand on regarde le modèle explicable, on s'aperçoit qu'en fait l'explication ne nous convient pas. L'explication montre surtout qu'il y avait un déséquilibre dans notre *dataset* d'apprentissage […] qui ne peut pas être à la base d'une réflexion biologique sur l'agressivité de la tumeur.
>
> — Pathologiste, entretien septembre 2022

En somme, il serait impossible au pathologiste de se prononcer sur la valeur de la prédiction en sortie d'un réseau.

Rendre les résultats des IA intelligibles : de l'explicabilité au « forçage » des modèles

La littérature en *XAI* offre deux principales stratégies pour résoudre le problème lié à l'opacité des réseaux de neurones convolutifs. La première recommande d'abandonner ces derniers au profit de modèles interprétables dits « par nature » ; la seconde recommande l'utilisation de techniques d'explication *post-hoc*. Ces deux stratégies offrent des avantages distincts, mais présentent également certaines limites, comme le montre le tableau ci-dessous.

Tableau 1 : Avantages et inconvénients des approches de l'*XAI*

Stratégie	Avantages	Inconvénients
Modèles interprétables	Résultats facilement interprétables (Rudin, 2019)	Performances réduites sur des données complexes
Techniques *post-hoc*	Permettent d'expliquer les résultats de modèles complexes (Holzinger *et al.*, 2022)	Risque d'explications inexactes (Mittelstadt *et al.*, 2019)

Les modèles interprétables facilitent la mise en relation des résultats avec les connaissances du praticien, mais perdent en efficacité à mesure que les données se complexifient. Les techniques *post-hoc*, bien qu'adaptées à ces

données, ne permettent pas d'établir ce lien de manière claire. De telles limites ont conduit le pathologiste et le *data scientist* impliqués dans APRIORICS à développer, par eux-mêmes, une solution permettant au pathologiste de donner du sens aux résultats produits par les modèles en les reliant à ses connaissances, y compris lorsque les données d'entrée sont complexes. Cette solution diffère très sensiblement des deux approches conventionnelles de l'*XAI*, en ceci qu'elle ne vise justement pas à rendre les mécanismes internes des modèles compréhensibles par le praticien. L'objectif est plutôt de s'assurer que les modèles ne puissent produire *que* des résultats intelligibles et utilisables par le praticien, sans pour autant exiger l'explicabilité des modèles eux-mêmes. Pour cela, la solution retenue dans APRIORICS s'est concentrée sur la structuration des données d'entrée, afin que les modèles, quel que soit leur fonctionnement interne, génèrent systématiquement des résultats pertinents pour le praticien.

Cette solution a consisté à contraindre les modèles de segmentation à localiser et à identifier correctement les objets biologiques présents dans les lames virtuelles. En d'autres termes, il s'agit de forcer les modèles à ne pas pouvoir faire autrement que fonder leurs prédictions sur les caractéristiques biologiquement pertinentes des données. Mais ce n'est pas en intervenant sur l'architecture des réseaux de neurones convolutifs que le *data scientist* impliqué dans le projet force ces derniers à segmenter correctement les objets biologiques présents dans l'image. Pour amener les modèles à générer des résultats biologiquement intelligibles, le projet APRIORICS a plutôt porté sur l'annotation des données d'entraînement et de validation des modèles. Après l'entraînement, il est attendu que les objets segmentés correspondent aux objets biologiques effectivement présents sur la lame et soient correctement étiquetés. C'est-à-dire que, par exemple, deux noyaux de cellules qui se chevauchent doivent être identifiés comme tels, et non comme une cellule en cours de mitose. Ainsi que le dit le pathologiste :

> Là, quand j'ai parlé de ce qu'il y avait dans une tumeur, j'ai dit « cellules tumorales », j'ai dit « cellules immunitaires », j'ai dit « vaisseaux », et les cellules sont constituées de « noyaux » et de « cytoplasmes », etc. Donc l'idée, c'est […] que la machine arrive à identifier ça.
>
> — Pathologiste, entretien septembre 2022

DES IA RENDUES INTELLIGIBLES PAR L'ANNOTATION DE LEURS DONNÉES D'ENTRAÎNEMENT

Une approche centrée sur la « vérité de terrain » des modèles de *deep learning*

Les modèles employés dans le cadre d'APRIORICS sont représentatifs de l'état de l'art dans le domaine de la segmentation d'images[16]. Accessibles depuis des plateformes dédiées telles que *GitHub* et *PyTorch Hub*, ces modèles ont été préalablement entraînés sur *ImageNet*, une vaste base de données d'images annotées. Ce pré-entraînement leur a permis d'acquérir la capacité à extraire et à utiliser des caractéristiques générales, utiles pour un large éventail de tâches de segmentation. Toutefois, *ImageNet* se compose principalement de photographies couleurs d'objets du quotidien, comme des avions, des voitures ou des animaux, capturées dans divers contextes et conditions d'éclairage (Deng *et al*., 2009). Ces images présentent des différences notables par rapport aux images histologiques. Par exemple, dans les images issues d'*ImageNet*, les objets sont généralement facilement discernables de leur environnement grâce à des contrastes prononcés de couleur, de texture ou de forme. À l'opposé, dans les images histologiques, les objets d'intérêt peuvent se confondre avec un environnement riche en structures similaires, ce qui exige une sensibilité accrue des modèles aux variations subtiles de l'image. Par conséquent, les modèles pré-entraînés sur *ImageNet* s'avèrent inadaptés à la segmentation d'objets biologiques dans des lames virtuelles. Leur inadéquation à la tâche est d'ailleurs renforcée par le fonctionnement intrinsèque des réseaux de neurones convolutifs :

> Plus on va profond dans le réseau, plus il va détecter des choses spécifiques aux sujets que l'on traite. Dans les premières couches, il va détecter peut-être des angles, des arrondis, des points. Et plus on va profondément, plus il va détecter des choses qui vont lui permettre de faire sa classification. Et donc si on récupère un réseau pré-entraîné, on imagine que les couches les plus profondes ne nous sont pas très utiles.
>
> — *Data scientist*, entretien mars 2023

[16] C'était le cas lorsque le projet a commencé, en 2021. Toutefois, sur certaines tâches, ces modèles sont aujourd'hui concurrencés par les *transformers*, un autre type d'architecture en apprentissage profond (Yao, 2024). Issus du domaine du traitement du langage naturel, ces modèles appréhendent l'image comme une séquence d'éléments (et non plus comme une grille de pixels, comme cela est le cas des réseaux de neurones convolutifs) à laquelle ils appliquent des mécanismes d'attention qui leur permettent d'y identifier des relations complexes, échappant aux réseaux de neurones convolutifs (Vaswani *et al*., 2017).

Ainsi, les modèles pré-entraînés nécessitent un ajustement aux exigences spécifiques de la tâche qui leur est assignée dans le cadre du projet. Ce processus d'ajustement, appelé « *transfer learning* », permet d'exploiter les caractéristiques générales déjà apprises par les modèles, tout en les adaptant pour que ces derniers puissent identifier des caractéristiques spécifiques aux images histologiques, telles que les formes distinctives des cellules tumorales[17]. Le *transfer learning* permet ainsi d'obtenir des modèles optimisés pour la segmentation d'objets biologiques, pour un coût computationnel et temporel bien moindre que celui nécessaire à la conception ou à l'entraînement de modèles *ex nihilo*[18]. En effet, comme nous allons le montrer, dans APRIORICS, l'effort pour contraindre les modèles à appréhender correctement les objets biologiques présents dans une lame a principalement porté sur la construction de ce que la communauté informatique appelle la « vérité de terrain » des modèles, et plus particulièrement sur l'annotation des données qu'elle contient :

> On n'avait pas prévu de faire de la recherche vraiment en *deep learning* en testant des architectures, en modifiant des architectures. L'idée, c'était quand même d'utiliser *a priori* des choses préexistantes, parce que l'énergie du *data scientist* avec la compétence d'analyse d'images devait être dirigée vers la partie traitement d'images du départ.
>
> — Pathologiste, entretien décembre 2022

La difficulté d'annoter manuellement des objets biologiques microscopiques

La « vérité de terrain » d'un modèle d'apprentissage automatique est l'ensemble de données utilisé à la fois pour entraîner et évaluer celui-ci. Florian Jaton a montré que ces vérités de terrain faisaient l'objet d'un travail de problématisation de la part des concepteurs des IA, orientant ainsi de façon décisive le processus d'apprentissage de celles-ci et, *in fine*, les résultats qu'elles génèrent (2017). Jaton suggère ainsi, déjà, que l'intelligibilité des résultats est prioritairement la résultante des choix faits lors de la constitution

[17] Le *transfer learning* consiste à réentraîner les couches profondes des modèles avec des images histologiques, tout en maintenant les premières couches inchangées – c'est-à-dire gelées.

[18] Au *transfer learning* s'ajoute un travail de « *fine-tuning* » (ajustement des hyperparamètres du modèle) et de « *post-processing* » (réorganisation du modèle, combinaison avec d'autres, etc.) permettant d'adapter plus finement les modèles à la tâche.

des vérités de terrain, en amont du traitement de ces données par les modèles (2022). Dès lors, rendre les résultats intelligibles passe moins par un effort pour rendre les modèles explicables, que par ce travail de problématisation effectué sur les données qui servent à les entraîner et à les valider.

Pour l'apprentissage mis en œuvre dans APRIORICS, ces vérités de terrain comprennent des données d'entrée, appariées à des annotations qui définissent la sortie attendue pour chacune d'entre elles. Durant la phase d'apprentissage, le modèle devra apprendre la fonction qui associe chaque entrée à l'annotation correspondante. La qualité et la diversité des données, ainsi que la précision des annotations contenues dans la vérité de terrain ont donc une influence significative sur l'exactitude des prédictions du modèle sur de nouvelles données (Oakden-Rayner, 2017).

Dès lors, apprendre à des réseaux de neurones convolutifs à segmenter correctement les objets biologiques présents dans une lame virtuelle implique de construire une vérité de terrain contenant des exemples de ces objets annotés avec précision. Par ailleurs, étant donné que la forme, la texture, et la couleur de ces objets peuvent varier au sein d'une même image et d'une image à l'autre en raison de facteurs liés à la préparation des lames, aux paramètres du scanner, et à certains états biologiques et pathologiques, cette vérité de terrain doit contenir des données diversifiées, qui reflètent cette variabilité. Tel qu'il est formulé, le problème qu'auront à résoudre les modèles implique donc d'annoter potentiellement des centaines de millions d'objets à l'échelle cellulaire.

En ACP, la tâche d'annotation est le plus souvent réalisée par les pathologistes, dans la mesure où seuls ceux-ci possèdent la compétence visuelle permettant de discerner les structures tissulaires et cellulaires[19]. Cette tâche consiste à identifier, puis à délimiter des structures biologiques d'intérêt, telles que des cellules en mitose ou des zones de tissu inflammatoire. La variabilité et la densité des images histologiques rendent ce travail particulièrement fastidieux : « c'est ingrat, c'est compliqué, c'est long », indique ainsi le pathologiste du projet étudié. En outre, bien que cette tâche repose principalement sur des critères biologiques objectifs, elle fait parfois appel à l'interprétation subjective :

[19] La tâche d'annotation des images histologiques peut être réalisée par des techniciens et scientifiques biomédicaux, voire par des logiciels à base d'apprentissage automatique, sous la supervision du pathologiste.

> Il y a pas mal de fois où l'on n'est pas vraiment sûrs de ce que l'on fait. Par exemple, quand il s'agit d'identifier les noyaux de cellules tumorales, dans 90 % des cas, je suis assez sûr de moi. Mais dans les 10 % restants, je doute. Est-ce que c'est vraiment une cellule tumorale ? Est-ce qu'il y a un ou deux noyaux ?
>
> — Pathologiste, entretien décembre 2022

Cette interprétation est compliquée par le fait que l'analyse de lames virtuelles induit une perte de détails par rapport à l'observation d'échantillons au microscope (Carboni *et al*., 2023). Par exemple, confronté à des structures qui se chevauchent dans l'image, le pathologiste ne peut ajuster la profondeur de champ ou la mise au point – comme il le ferait avec un microscope – pour mieux distinguer ces structures.

En raison de la difficulté du travail d'annotation, les vérités de terrain disponibles pour la segmentation d'images histologiques sont de deux types. Le premier comprend des vérités de terrain contenant un grand nombre de lames virtuelles, annotées à une échelle relativement générale. Par exemple, l'annotation peut se limiter à indiquer la présence d'une tumeur dans l'image, sans fournir de détails au niveau cellulaire. Le second type regroupe des vérités de terrain composées en moyenne d'une centaine de *patches*, c'est-à-dire de portions de lames virtuelles, qui sont annotés – avec une précision variable – à l'échelle cellulaire[20]. Bien que les vérités de terrain de ce type soient adaptées à des tâches de segmentation telles que celle envisagée dans APRIORICS, la quantité limitée de données qu'elles contiennent et la qualité de leurs annotations peuvent affecter les performances des modèles développés. Par conséquent, dans un cas comme dans l'autre, les vérités de terrain actuellement disponibles se révèlent insuffisantes pour atteindre les objectifs fixés dans le cadre du projet étudié.

L'immunohistochimie au service d'une stratégie d'annotation automatique

L'immunohistochimie détournée

[20] Une lame virtuelle mesure généralement 100 000 x 100 000 pixels. Imprimée à pleine résolution avec une excellente qualité d'impression (300 dpi), elle mesurerait plus de 3 m de côté. Un *patch* mesure en général 200 x 200 pixels.

Pour surmonter les limitations imposées par l'annotation manuelle au domaine de la segmentation d'images histologiques, l'équipe a opté pour l'automatisation de cette tâche.

Dans certains domaines, comme le montre Camille Girard-Chanudet à propos du travail d'annotation de décisions de justice, les annotatrices et annotateurs ont la possibilité de créer des « cadres catégoriels intermédiaires » pour y faire entrer des cas se laissant difficilement subsumer sous les catégories disponibles (2023). Le pathologiste, quant à lui, n'a pas la liberté d'adopter cette même stratégie. Cela le conduirait en effet à « créer » des objets biologiques nouveaux – l'équivalent des cadres catégoriels intermédiaires de Girard-Chanudet –, ce qui serait contre-productif, voire dangereux. Dans le premier cas, le travail d'annotation fait intervenir ce qu'Emmanuel Kant appelait un « jugement réfléchissant », au sens où le cas donne la règle ; dans le second cas, au contraire, la règle est donnée et le cas y est subsumé : il s'agit de ce que Kant appelait un « jugement déterminant ». Cette différence entre deux formes de jugement est importante, parce qu'elle conditionne la possibilité ou non d'automatiser la tâche d'annotation. En effet, le jugement réfléchissant tire la règle du cas, ce qui implique un travail d'interprétation et de contextualisation qui se laisse difficilement automatiser. Le jugement déterminant, quant à lui, applique machinalement la règle au cas, ce qui, de ce fait, permet d'envisager son automatisation.

Le choix original adopté par le pathologiste et le *data-scientist* pour automatiser l'annotation des vérités de terrain a consisté à recourir à une technique bien connue en ACP et utilisée en routine clinique par le pathologiste : l'immunohistochimie[21]. Celle-ci permet de détecter des protéines spécifiques à un type, une fonction ou une structure cellulaire au moyen d'une réaction antigène-anticorps. Elle est couramment employée pour préciser un diagnostic, établir un pronostic et prédire la réponse à certains traitements en mesurant l'expression de marqueurs caractéristiques de certains événements cellulaires. Dans le cadre du projet, cette technique est détournée de sa fonction principale pour marquer les différents composants de la tumeur puis, à partir de ce marquage, créer des masques de segmentation binaires qui

[21] Quelques équipes dans le monde ont élaboré des stratégies d'annotation analogues (Kataria *et al*., 2023 ; Lin *et al*., 2023). Cependant, le projet APRIORICS se démarque de ces expérimentations par la taille de l'ensemble de données qu'il a permis de produire et la diversité des structures biologiques qui ont été annotées.

servent d'annotations en spécifiant, dans l'image, quels pixels appartiennent à l'objet d'intérêt et quels pixels relèvent de son arrière-plan[22].

Des prédictions ancrées dans la réalité biologique

L'obtention de ces masques procède comme suit : une cohorte de 1250 patientes, reflétant l'hétérogénéité des tumeurs du sein et de leurs approches thérapeutiques, est constituée. Pour chaque patiente, un bloc de paraffine contenant un échantillon représentatif de la tumeur est sélectionné. La fixation dans le formol et l'inclusion en paraffine visent à préserver la morphologie des structures tissulaires et cellulaires et à permettre la réalisation de coupes fines. À partir de chaque bloc, neuf lames en coloration standard sont préparées – ce chiffre correspondant aux différents composants de la tumeur examinés dans le projet et représentant les « briques élémentaires » du cancer du sein. Après avoir numérisé ces lames, un protocole d'immunohistochimie est appliqué sur les mêmes coupes tissulaires pour marquer, dans chacune d'elles, un composant tumoral particulier. À l'issue de ce marquage, ces coupes sont montées sur de nouvelles lames, qui sont numérisées à leur tour.

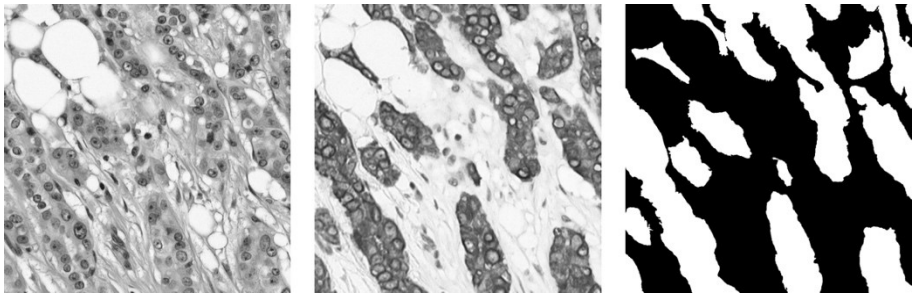
Cette stratégie permet donc d'obtenir neuf paires de lames virtuelles pour chaque patiente : neuf images en coloration standard, chacune appariée à une image immunohistochimique qui marque un élément spécifique de la tumeur. Toutefois, étant donné la taille considérable de ces images, de l'ordre de 100 000 x 100 000 pixels, avec un poids approximatif de 2 Go chacune, il est nécessaire de découper celles-ci en *patches* plus petits.

Ce découpage ne sert pas uniquement à réduire la taille des images pour permettre à un modèle d'apprentissage automatique de les traiter ; il contribue également à augmenter de façon significative la quantité de données disponibles pour constituer la vérité de terrain. Ainsi, « on peut extraire 4,5 millions de *patches* à partir de 500 lames », explique le *data scientist*. Des techniques de traitement d'images sont ensuite appliquées aux *patches* issus des images immunohistochimiques afin d'en extraire des masques de segmentation binaires. Ces masques, qui sont parfaitement superposables aux

[22] Plus précisément, un masque de segmentation binaire est une image qui en simplifie une autre en réduisant les informations que contient cette dernière à deux valeurs distinctes pour chaque pixel : 1 (ou blanc) pour indiquer que le pixel appartient à l'objet d'intérêt, et 0 (ou noir) pour signaler que le pixel ne fait pas partie de cet objet.

patches en coloration standard, délimitent précisément les objets biologiques que les réseaux de neurones convolutifs doivent apprendre à segmenter.

Figure 2. Triplet de *patches* pour un même échantillon de tissu



Source : Dr Camille Franchet et Robin Schwob.

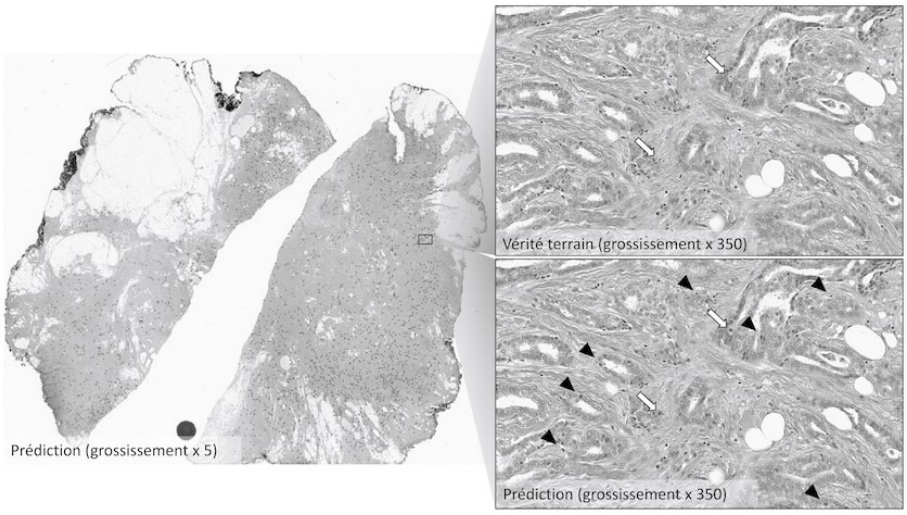
De gauche à droite : image en coloration standard (HE), immunomarquage avec l'anticorps anti-AE1/AE3 (marquage des cellules épithéliales), masque binaire des cellules épithéliales.

Ainsi, ce protocole d'annotation automatique permet non seulement à l'équipe d'obtenir une vérité de terrain bien plus étendue et précisément annotée que l'état de l'art dans de nombreux domaines de la segmentation d'images histologiques, mais il garantit également, *a priori*, l'exactitude des prédictions des modèles entraînés à partir de ces données. En effet, l'immunohistochimie, reconnue pour sa précision et sa spécificité dans la détection de protéines, produit des annotations qui reflètent fidèlement les objets biologiques à segmenter. Ce faisant, elle confère aux prédictions des modèles un fondement biologique solide, rendant superflue l'explication de leur fonctionnement interne.

Bien que la démarche adoptée dans APRIORICS puisse sembler s'inscrire dans une tendance générale privilégiant la performance prédictive des modèles au détriment de leur compréhension (Vuarin et Steyer, 2023), en réalité, cette stratégie renforce l'intelligibilité des résultats des modèles. En particulier, l'analyse de la vérité de terrain révèle les raisons sous-jacentes aux prédictions de ceux-ci et, si nécessaire, permet de réaliser des ajustements ciblés pour améliorer leur exactitude. Par exemple, lors de l'un de nos entretiens, le *data scientist* indique qu'un modèle conçu pour segmenter des mitoses, qui sont des objets biologiques relativement rares sur une lame

histologique, tend à en segmenter trop (900, au lieu des 19 réellement présentes dans l'image).

Figure 3. Segmentation des mitoses sur une image microscopique de cancer du sein en coloration standard



Source : Dr Camille Franchet

À gauche, la prédiction du modèle sur l'ensemble de l'image. En haut à droite, la vérité de terrain, montrant les mitoses réellement présentes dans une portion de l'image (indiquées par des flèches blanches). En bas à droite, les mitoses prédites par le modèle dans la même zone. Les vrais positifs sont désignés par des flèches blanches ; les faux positifs par des têtes de flèches noires.

L'examen de la vérité de terrain révèle que le modèle peine à reconnaître la véritable forme d'une mitose sur la base des annotations fournies, car ces dernières ne capturent pas avec exactitude les critères morphologiques permettant de différencier les mitoses de structures cellulaires similaires, telles que les noyaux de certaines cellules. Ce constat ouvre alors plusieurs pistes pour diminuer le nombre de faux positifs : « est-ce qu'il faut qu'on affine notre vérité de terrain pour être plus précis ? Est-ce qu'il faut qu'on fasse du *post-processing* avec de la connaissance biologique ? », s'interroge le *data-scientist*.

La généralisabilité des modèles, autre aspect essentiel à l'intégration des IA en médecine

Ainsi, en produisant des annotations très précises, la stratégie développée dans le cadre du projet APRIORICS permet aux réseaux de neurones convolutifs d'apprendre les caractéristiques des données distinctives des objets biologiques à segmenter. Cette approche contraint les modèles à fonder leurs prédictions sur des objets biologiques réels. Bien que cela ne rende pas leur fonctionnement interne plus explicable pour les médecins, les résultats en sortie possèdent nécessairement une pertinence biologique, ce qui leur donne du sens aux yeux des praticiens. L'intelligibilité de ces résultats est donc opératoire, même si les modèles eux-mêmes restent opaques.

Pourtant, cette approche n'assure pas entièrement la généralisabilité des modèles, c'est-à-dire leur capacité à reconnaître correctement ces objets dans de nouvelles images, y compris celles provenant d'autres laboratoires d'ACP. Le fait est qu'une partie des caractéristiques apprises par les modèles peuvent être liées au contexte d'apprentissage, une problématique nommée « sur-apprentissage ».

En effet, l'essor de la pathologie digitale s'est accompagné d'un effort de standardisation des conditions de production des lames de verre, et ce, afin d'assurer qu'une fois numérisées, celles-ci permettent une qualité d'interprétation comparable à l'observation au microscope. Toutefois, cette standardisation n'a pas complètement éliminé la variabilité des pratiques en ACP. Ainsi, les lames virtuelles produites par différents laboratoires présentent des variations, attribuables notamment à des différences dans les protocoles de préparation des échantillons de tissus, dans l'équipement de numérisation, et dans la sélection des cas. Par exemple, parmi les techniques de coloration de routine des échantillons tissulaires, la coloration HE, qui teint les noyaux en bleu et les cytoplasmes et la fibrose en rose, est la plus répandue à travers le monde. Cependant, les laboratoires d'ACP français font exception, en privilégiant encore souvent la coloration hématoxyline-éosine+safran (HE+S), où le safran ajoute aux nuances de bleu et de rose du jaune, qui permet une meilleure distinction de certains tissus conjonctifs. Ainsi, les lames virtuelles obtenues à partir de lames en coloration HE et HE+S présentent des différences visuelles notables, particulièrement en termes de contraste et de saturation des couleurs. L'œil du pathologiste est capable de surmonter ces différences, car ces deux méthodes de coloration révèlent les mêmes structures cellulaires et tissulaires. En revanche, pour un réseau de neurones convolutif, qui analyse les images en se fondant uniquement sur des

valeurs de pixels, ces images apparaîtront comme différentes. Par conséquent, un réseau entraîné sur des lames HE+S numérisées aura des difficultés à identifier les caractéristiques des données apprises dans des lames HE numérisées.

Ainsi, en raison de ces variations inter-laboratoires, les modèles entraînés à partir de la vérité de terrain construite dans le cadre du projet APRIORICS pourraient afficher des performances sous-optimales dans d'autres laboratoires. Or, d'après le pathologiste :
« il y avait un tel travail technique pour produire la donnée annotée qu'on pouvait difficilement proposer une approche multicentrique, avec plusieurs centres qui auraient fait la même chose, pour capter un peu mieux cette variabilité inter-laboratoires [dans la vérité de terrain] ».

En effet, un investissement considérable en temps, personnel et équipement a été nécessaire pour préparer plus de 9000 lames, les démonter sans endommager les échantillons de tissus qu'elles renferment, appliquer différents protocoles d'immunohistochimie sur ces échantillons, numériser l'ensemble, corriger manuellement les distorsions liées aux techniques de préparation et de numérisation sur les lames virtuelles, extraire des masques de ces images, et enfin, ajuster ces masques pour qu'ils correspondent précisément aux objets biologiques qu'ils sont censés recouvrir. Réaliser un travail d'une telle envergure dans d'autres laboratoires s'avère particulièrement difficile. Ainsi, quand bien même les modèles développés dans le cadre d'APRIORICS présenteraient de bonnes performances à l'IUCT-O, rien ne permet d'affirmer qu'il en serait de même dans d'autres laboratoires d'ACP.

Pourtant, l'ambition du projet est précisément d'améliorer la capacité d'analyse des pathologistes, afin de réaffirmer la centralité de ces spécialistes dans la prise en charge des cancers. En ce sens, les modèles développés dans le cadre d'APRIORICS ont vocation à être largement utilisés. Dans ces conditions, la solution consiste notamment à faire de l' « augmentation de données », c'est-à-dire à ajuster aléatoirement les couleurs des images, afin d'obtenir une vérité de terrain représentative des pratiques d'autres laboratoires (Franchet *et al*., 2024). Une autre solution pourrait être de faire de l'« adaptation de domaine », une technique qui permettrait aux modèles de reconnaître les caractéristiques des données apprises sur les images d'autres laboratoires à partir d'un petit ensemble de données (annotées ou non). La *startup* israélienne Ibex, qui commercialise un assistant pour la lecture de biopsies de prostates, utilise cette solution. Selon le pathologiste : « si nous

voulons utiliser l'outil, ils vont nous dire : ''donnez-nous quarante lames de biopsies de prostate de votre laboratoire pour que l'on fasse une adaptation de domaine'' ». Cependant, le *data scientist* note que pour le moment, cette solution « n'est pas forcément hyper documentée. Il y a des solutions brevetées ».

En somme, l'accent mis, dans la littérature, sur la question de l'explicabilité des IA médicales ne doit pas éclipser l'importance d'aspects tout aussi essentiels pour la sécurité des patients et la confiance des professionnels de santé, tels que la généralisabilité des modèles. Or, ce cas, qui illustre un problème fréquemment rencontré par les *data scientists* en ACP, montre que garantir cette généralisabilité dès la phase de conception peut s'avérer difficile à mettre en œuvre (Stacke *et al*., 2023).

CONCLUSION

Cet article met en évidence le décalage, déjà souligné dans la littérature, entre l'explicabilité des IA telle qu'elle est généralement envisagée dans le champ de l'*XAI*, au sens de la compréhension de leurs mécanismes internes, et l'intelligibilité des résultats produits par les modèles, telle qu'elle est conçue et souhaitée par les utilisateurs (Lipton, 2017b). Contrairement à une idée très répandue, l'explicabilité des IA n'est pas indispensable pour atteindre cette intelligibilité, et donc pour l'appropriation de ces technologies par les médecins. Ainsi, dans le cas que cet article a pris pour objet, plutôt que de rendre le fonctionnement des modèles explicable au sens de l'*XAI*, il s'est plutôt agi de créer une solution d'annotation forçant les modèles à fonder leurs prédictions sur les caractéristiques biologiquement pertinentes des données. Les résultats se trouvent alors alignés avec les connaissances du pathologiste, ce qui lui permet de les comprendre, de les évaluer et de les intégrer efficacement dans sa pratique.

En ce sens, l'analyse indique que la question de l'appropriation des IA doit être posée en faisant grand cas des spécificités de chaque contexte d'usage, et des préoccupations des utilisateurs finaux. Cela requiert une collaboration étroite et continue entre les concepteurs et les utilisateurs. Dans le cadre du projet APRIORICS, cette collaboration a été particulièrement active, mais il convient de souligner que ce niveau d'implication des praticiens dans la conception de dispositifs à base d'IA n'est pas toujours atteint dans tous les projets. Cela étant dit, on peut supposer qu'une des conditions du succès des

IA médicales réside dans la capacité des concepteurs à intégrer, d'une manière ou d'une autre, les connaissances du médecin dans ces dispositifs. À ce jour, seules quelques initiatives, comme *Sepsis Watch*™, un logiciel de prédiction du sepsis, ont adopté cette approche contextuelle de la conception d'IA (Sendak *et al*., 2019).

L'*XAI* manifeste certes une volonté croissante de combler ce décalage. Cette tendance se traduit par des efforts pour caractériser finement les attentes des utilisateurs au moyen d'enquêtes auprès de panels, dans le but d'identifier, parmi les méthodes explicatives existantes, celles qui sont appropriées à ces besoins (Doshi-Velez et Kim, 2017 ; Tonekaboni *et al*., 2019 ; Arbelaez Ossa *et al*., 2022). Ces enquêtes sont utiles dans la mesure où l'explicabilité des IA peut correspondre à un réel besoin de certains utilisateurs. L'article n'a aucunement voulu démontrer le contraire, mais simplement que l'explicabilité ne pouvait pas *systématiquement* être tenue pour une condition *sine qua non* de l'intelligibilité, par les utilisateurs, des prédictions faites par les modèles, et donc de leur appropriabilité par ces derniers. Cela étant dit, même quand il est attesté que les utilisateurs ont effectivement un besoin d'explicabilité des IA, leurs attentes en la matière peuvent varier au sein d'une même spécialité médicale, selon l'expérience individuelle, ou encore selon les spécificités de leurs tâches, qui peuvent différer significativement (Evans *et al*., 2022). Or, l'*XAI* tend à négliger cette diversité des besoins en matière d'explicabilité. À supposer qu'elle soit le problème à résoudre, celle-ci est aussi dépendante des contextes.

L'étude d'APRIORICS a en outre fait émerger trois autres considérations importantes, qui pourront donner lieu à des recherches plus approfondies : premièrement, les IA peuvent se heurter à la variabilité des conditions des « milieux » dans lesquels elles sont développées, comme le montre l'exemple de la diversité des colorants utilisés pour préparer les lames de verre. Dans le projet APRIORICS, l'automatisation de l'annotation, si elle permet de s'assurer de l'exactitude des prédictions, n'assure pas pour autant la validité externe des modèles et donc la possibilité de les exporter dans d'autres laboratoires. En effet, cela exigerait une homogénéisation des lames virtuelles, et donc une standardisation des processus qui permettent de produire celles-ci, ce qui actuellement n'est pas le cas. Les IA n'ont pas de pouvoir « d'agentivité » en elles-mêmes : leur « agentivité » ne peut s'exercer que dans un milieu qui présente certaines propriétés spécifiques, essentielles pour que les IA produisent les effets escomptés (Verbeek, 2005). Cela invite à repositionner la question des enjeux des intelligences artificielles dans les contextes sociotechniques où elles s'insèrent.

Deuxièmement, cette étude montre que l'IA ne remplace ni n'affaiblit inévitablement l'expertise du praticien. Alors même que les études existantes sur l'IA en ACP indiquent que celle-ci induit une perte par rapport aux méthodes traditionnelles du pathologiste, le projet APRIORICS montre que l'IA peut au contraire représenter un moyen, pour le praticien, de renforcer son expertise face à des expertises concurrentes (Procter *et al*., 2024). Cela conduit à s'intéresser à la manière dont l'IA reconfigure cette expertise et la repositionne dans la prise en charge des cancers.

Troisièmement, APRIORICS propose une stratégie originale pour associer les approches statistiques propres aux réseaux de neurones profonds et les connaissances de l'expert, sans pour autant que cette association ne relève des approches neuro-symboliques, qui tentent de combiner approches statistiques et symboliques, notamment en formalisant les connaissances de l'expert sous forme de règles.

RÉFÉRENCES

AMANN J., BLASIMME A., VAYENA E., FREY D., MADAI V. (2020), Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, *BMC Medical Informatics and Decision Making*, vol. 20, p. 1-9. DOI: 10.1186/s12911-020-01332-6

AMGAD M., HODGE J. M., ELSEBAIE M. A. T., BODELON C., PUVANESARAJAH S., GUTMAN D. A., SIZIOPIKOU K. P., GOLDSTEIN J. A., GAUDET M. M., TERAS L. R., COOPER L. A. D. (2024). A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nature Medicine*, vol. 30, n°1, p. 85-97. DOI: 10.1038/s41591-023-02643-7

ARBELAEZ OSSA L., STARKE G., LORENZINI G., VOGT J. E., SHAW D. M., ELGER B. S. (2022), Re-focusing explainability in medicine, *Digital Health*, vol. 8, p. 1-9. DOI: 10.1177/20552076221074488

BARREDO ARRIETA A., DÍAZ-RODRÍGUEZ N., DEL SER J., BENNETOT A., TABIK S., BARBADO A., GARCIA S., GIL-LOPEZ S., MOLINA D., BENJAMINS R., CHATILA R., HERRERA F. (2020), Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, vol. 58, p. 82-115. DOI: 10.1016/j.inffus.2019.12.012

BERGERON H., CASTEL P., VÉZIAN A. (2021), Beyond full jurisdiction: Pathology and inter-professional relations in precision medicine, *New Genetics and Society*, vol. 40, n° 1, p. 42-57. DOI: 10.1080/14636778.2020.1861543

BOURRET P., KEATING P., CAMBROSIO A. (2011), Regulating diagnosis in post-genomic medicine: Re-aligning clinical judgment?, *Social Science & Medicine*, vol. 73, n° 6, p. 816-824. DOI: 10.1016/j.socscimed.2011.04.022

CARBONI C., WEHRENS R., VAN DER VEEN R., DE BONT A. (2023), Eye for an AI: More-than-seeing, fauxtomation, and the enactment of uncertain data in digital pathology, *Social Studies of Science*, vol. 53, n° 5, p. 712-737. DOI: 10.1177/03063127231167589

CARDON D., COINTET J.-P., MAZIÈRES A. (2018), La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle, *Réseaux*, vol. 211, n° 5, p. 173-220. DOI: 10.3917/res.211.0173

CCNE et CNPEN (2022), *Diagnostic médical et intelligence artificielle : enjeux éthiques*, Avis commun du CCNE et du CNPEN, Avis 141 du CCNE, Avis 4 du CNPEN.

DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L. (2009), ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami (États-Unis), 20-25 juin.

DOSHI-VELEZ F., KIM B. (2017), *Towards a rigorous science of interpretable machine learning*, arXiv. DOI: 10.48550/arXiv.1702.08608

DOSHI-VELEZ F., KORTZ M., BUDISH R., BAVITZ C., GERSHMAN S., O'BRIEN D., SCOTT K., SCHIEBER S., WALDO J., WEINBERGER D., WELLER A., WOOD A. (2019), *Accountability of AI Under the Law: The role of explanation*, décembre, Harvard University, Harvard Law School, Berkman Klein Center for Internet & Society, Cambridge, États-Unis, [En ligne] Disponible à l'adresse : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761, (consulté le 01/04/2024).

DROGT J., MILOTA M., VOS S., BREDENOORD A., JONGSMA K. (2022), Integrating artificial intelligence in pathology: A qualitative interview study of users' experiences and expectations, *Modern Pathology*, vol. 35, n° 11, p. 1540-1550. DOI: 10.1038/s41379-022-01123-6

EVANS T., RETZLAFF C. O., GEIßLER C., KARGL M., PLASS M., MÜLLER H., KIEHL T.-R., ZERBE N., HOLZINGER A. (2022), The explainability paradox: Challenges for xAI in digital pathology, *Future Generation Computer Systems*, vol. 133, p. 281-296. DOI: 10.1016/j.future.2022.03.009

FRANCHET C., SCHWOB R., BATAILLON G., SYRYKH C., PERICART S., FRENOIS F.-X., PENAULT-LLORCA F., LACROIX-TRIKI M., ARNOULD L., LEMONNIER J., ALLIOT J.-M., FILLERON T., BROUSSET P. (2024), Bias reduction using combined stain normalization and augmentation for AI-based classification of histological images, *Computers in Biology and Medicine*, vol. 171, p. 1-9. DOI: 10.1016/j.compbiomed.2024.108130

GAGLIO G., LOUTE A. (2023), L'émergence d'enjeux éthiques lors d'expérimentations de logiciels d'intelligence artificielle. Le cas de la radiologie, *Réseaux*, vol. 240, n° 4, p. 145-178. DOI: 10.3917/res.240.0145

GERDES A. (2024), The role of explainability in AI-supported medical decision-making, *Discover Artificial Intelligence*, vol. 4, p. 1-29. DOI: 10.1007/s44163-024-00119-2

GHASSEMI M., OAKDEN-RAYNER L., BEAM A. L. (2021), The false hope of current approaches to explainable artificial intelligence in health care, *The Lancet Digital Health*, vol. 3, n° 11, p. 745-750. DOI: 10.1016/S2589-7500(21)00208-9

GIRARD-CHANUDET C. (2023), « Mais l'algo, là, il va mimer nos erreurs ! » : Contraintes et effets de l'annotation des données d'entraînement d'une IA, *Réseaux*, vol. 240, n° 4, p. 111-144. DOI: 10.3917/res.240.0111

GUNNING D. (2017), « Explainable artificial intelligence (xAI) », Defense Advanced Research Projects Agency (site), mai.

HANNA M. G., PANTANOWITZ L. (2017), Why is digital pathology in cytopathology lagging behind surgical pathology?, *Cancer Cytopathology*, vol. 125, n° 7, p. 519-520. DOI: 10.1002/cncy.21855

HERZOG C. (2022), On the ethical and epistemological utility of explicable AI in medicine, *Philosophy & Technology*, vol. 35, n° 2, p. 1-31. DOI: 10.1007/s13347-022-00546-y

HOLZINGER A., SARANTI A., MOLNAR C., BIECEK P., SAMEK W. (2022), « Explainable AI methods: A brief overview », *in* HOLZINGER A. GOEBEL R., FONG, MOON T., MÜLLER K.-R., SAMEK W. (eds.), *xxAI: Beyond explainable AI*, Cham, Springer International Publishing, p. 13-38. [En ligne] Disponible à l'adresse : https://link.springer.com/chapter/10.1007/978-3-031-04083-2_1, (consulté le 01/04/2024).

JATON F. (2017), We get the algorithms of our ground truth: Designing referential databases in digital image processing, *Social Studies of Science*, vol. 47, n° 6, p. 811-840. DOI: 10.1177/0306312717730428

JATON F. (2022), De la centralité des bases de données *ground truth* dans la construction des systèmes algorithmiques, *Silomag*, juillet.

KATARIA T., RAJAMANI S., AYUBI A. B., BRONNER M., JEDRZKIEWICZ J., KNUDSEN B.S., ELHABIAN S. Y. (2023), Automating ground truth annotations for gland segmentation through immunohistochemistry, *Modern Pathology*, vol. 36, n° 12. DOI: 10.1016/j.modpat.2023.100331

LEBOVITZ S., LIFSHITZ-ASSAF H., LEVINA N. (2022), To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis, *Organization Science*, vol. 33, n° 1, p. 126-148. DOI: 10.1287/orsc.2021.1549

LIN T-P., YANG C.-Y., LIU K.-J., HUANG M.-Y., CHEN Y.-L. (2023), Immunohistochemical stain-aided annotation accelerates machine learning and deep learning model development in the pathologic diagnosis of nasopharyngeal carcinoma, *Diagnostics*, vol. 13, n° 24. DOI: 10.3390/diagnostics13243685

LIPTON Z. C. (2017a), *The doctor just won't accept that!,* arXiv. DOI: 10.48550/arXiv.1711.08037

LIPTON Z. C. (2017b), *The mythos of model interpretability,* arXiv. DOI: 10.48550/arXiv.1606.03490

MATEESCU A., ELISH M. C. (2019), *AI in context: the labor of integrating new technologies*, Data & Society Research Institute, New-York, États-Unis, [En ligne] Disponible à l'adresse : https://datasociety.net/wp-content/uploads/2019/01/DataandSociety_AIinContext.pdf, (consulté le 22/06/2024)

MILLER T., HOWE P., SONENBERG L. (2017), *Explainable AI: Beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences*, arXiv. DOI: 10.48550/arXiv.1712.00547

MINISTERE DU TRAVAIL, DE LA SANTE ET DES SOLIDARITES (2023, 16 juin), *François BRAUN lance une mission sur la pathologie digitale*. Ministère du Travail, de la Santé et des Solidarités. https://sante.gouv.fr/actualites/presse/communiques-de-presse/article/francois-braun-lance-une-mission-sur-la-pathologie-digitale.

MITTELSTADT B., RUSSELL C., WACHTER, S. (2019), Explaining explanations in AI, *in* Association for Computing Machinery (ed), *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New-York, Association for Computing Machinery, p. 279-288. [En ligne] Disponible à l'adresse : https://dl.acm.org/doi/10.1145/3287560.3287574, (consulté le 01/04/2023).

OAKDEN-RAYNER L. (2017), « Exploring the ChestXray14 dataset: Problems », OAKDEN-RAYNER L. (blog), 18 décembre.

PANTANOWITZ L., SHARMA A., CARTER A. B., KURC T., SUSSMAN A., SALTZ J. (2018), Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives, *Journal of Pathology Informatics*, vol. 9, n° 1, p. 1-40. DOI: 10.4103/jpi.jpi_69_18

PROCTER R., ROUNCEFIELD M., TOLMIE P., VERRILL C. (2024), Everyday diagnostic work in the histopathology lab: CSCW perspectives on the utilization of data-driven clinical decision support systems, *Computer Supported Cooperative Work (CSCW)*. DOI: 10.1007/s10606-024-09496-9

RATTI E., GRAVES M. (2022), Explainable machine learning practices: Opening another black box for reliable medical AI, *AI and Ethics*, vol. 2, n° 4, p. 801-814. DOI: 10.1007/s43681-022-00141-z

RUDIN C. (2019), Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, vol. 5, n° 1, p. 206-215. DOI: 10.1038/s42256-019-0048-x

SENDAK M., ELISH M. C., GAO M., FUTOMA J., RATLIFF W., NICHOLS M., BEDORA A., BALU S., O'BRIEN C. (2020), « "The human body is a black box": Supporting clinical decision-making with deep learning », *in* Association for Computing Machinery (ed), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New-York, Association for Computing Machinery, p. 99-109.

SIGAUT, F. (2012), *Comment homo devint faber : Comment l'outil fit l'homme*, Paris, CNRS Éditions.

STACKE K., EILERTSEN G., UNGER J., LUNDSTRÖM C. (2019), *A closer look at domain shift for deep learning in histopathology,* arXiv. DOI: 10.48550/arXiv.1909.11575

TONEKABONI S., JOSHI S., MCCRADDEN M. D., GOLDENBERG A. (2019), *What clinicians want: Contextualizing explainable machine learning for clinical end use*, arXiv. DOI: 10.48550/arXiv.1905.05134

VAN DER LAAK J., LITJENS G., CIOMPI F. (2021), Deep learning in histopathology: The path to the clinic, *Nature Medicine*, vol. 27, n° 5, p. 775-784. DOI: 10.1038/s41591-021-01343-4

VAN DIEST P. J., VAN DER WALL E., BAAK J. P. A. (2004), Prognostic value of proliferation in invasive breast cancer: A review, *Journal of Clinical Pathology*, vol. 57, n° 7, p. 675-681. DOI: 10.1136/jcp.2003.010777

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I. (2023), *Attention is all you need*, arXiv. DOI: 10.48550/arXiv.1706.03762

VERBEEK P.-P. (2005), *What things do: Philosophical reflections on technology, agency, and design*, University Park, Pennsylvania State University Press.

VERBEEK P.-P. (2010), Accompanying technology: Philosophy of technology after the ethical turn, *Techné: Research in Philosophy and Technology*, vol. 14, n° 1, p. 49-54. DOI: 10.5840/techne20101417

VERBEEK P.-P. (2022), « The empirical turn », *in* VALLOR S. (ed), *The Oxford Handbook of Philosophy of Technology*, Oxford, Oxford University Press, p. 35-54.

VUARIN L., STEYER, V. (2023), Le principe d'explicabilité de l'IA et son application dans les organisations, *Réseaux*, vol. 240, n° 4, p. 179-210. DOI: 10.3917/res.240.0179

WANG S., YANG D. M., RONG R., ZHAN X., XIAO G. (2019), Pathology image analysis using segmentation deep learning algorithms, *The American Journal of Pathology*, vol. 189, n° 9, p. 1686-1698. DOI: 10.1016/j.ajpath.2019.05.007

YAO W., BAI J., LIAO W., CHEN Y., LIU M., XIE Y. (2024), *From CNN to transformer: A review of medical image segmentation models*, arXiv. DOI: 10.48550/arXiv.2308.05305

CONCLUSION

Par son impact sociétal majeur, l'IA soulève d'importantes questions éthiques, juridiques et philosophiques. En médecine en général, et en pathologie en particulier, l'avènement de modèles d'IA d'assistance au diagnostic, ou d'évaluation du pronostic, suscite espoirs et craintes, parfois amplifiés par une forme de surenchère médiatique. Même si nous nous sentons à l'abri d'un 3ème hiver de l'IA, la mode actuelle va finir par décroître et l'IA prendra progressivement une place dans nos pratiques et dans nos vies, en suivant la courbe de la hype de Gartner. Au-delà de ces ressentis très subjectifs, la rigueur scientifique doit rester au centre des préoccupations et l'interaction étroite entre data scientists et médecins est au cœur de cette notion d'intégrité scientifique. Dans le cadre de ce projet de thèse, nous avons pu apprécier à quel point, en toute bonne foi, on pouvait considérer à tort qu'un modèle d'IA était performant. Ce fut l'occasion de découvrir le vaste champ des biais qui couvre à la fois des questions de société et des éléments plus spécifiques aux méthodes d'apprentissage automatique. La recherche, l'identification et le monitoring des biais dans les projets d'IA représentent un effort réel mais indispensable à une utilisation future, apaisée, d'outils d'IA dans notre pratique quotidienne. Il y a fort à parier que l'interprétabilité des modèles d'IA nourrira les réflexions médicales, éthiques, juridiques et philosophiques associées à ces nouveaux outils. Dans le domaine de l'interprétabilité, les méthodes basées sur le *feature engineering* et l'extraction de *features* connaissent un renouveau très prometteur dans lequel s'inscrit le projet APRIORICS.

V. REFERENCES BIBLIOGRAPHIQUES

[1] Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet 2010;11:855–66.

[2] Harder N, Athelogou M, Hessel H, et al. Tissue Phenomics for prognostic biomarker discovery in low- and intermediate-risk prostate cancer. Sci Rep 2018;8:4470.

[3] Beck AH, Sangoi AR, Leung S, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3:108ra113.

[4] MacGrogan G, Mathieu M-C, Poulet B, et al. [Pre-analytical stage for biomarker assessment in breast cancer: 2014 update of the GEFPICS' guidelines in France]. Ann Pathol 2014;34:366–72.

[5] Marck V. Manuel de techniques d'anatomo-cytopathologie - Théorie et pratique. Elsevier Masson. Elsevier Masson; 2010.

[6] Chan JKC. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. Int J Surg Pathol 2014;22:12–32.

[7] Titford M. The long history of hematoxylin. Biotech Histochem 2005;80:73–8.

[8] Titford M. Progress in the Development of Microscopical Techniques for Diagnostic Pathology. Journal of Histotechnology 2009;32:9–19.

[9] Travis AS. "Ambitious and Glory Hunting . . . Impractical and Fantastic": Heinrich Caro at BASF. Technology and Culture 1998;39:105.

[10] Weinstein RS, Graham AR, Richter LC, et al. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. Hum Pathol 2009;40:1057–69.

[11] Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. Histopathology 2012;61:1–9.

[12] Snead DRJ, Tsang Y-W, Meskiri A, et al. Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 2016;68:1063–72.

[13] Goacher E, Randell R, Williams B, Treanor D. The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review. Arch Pathol Lab Med 2017;141:151–61.

[14] Mukhopadhyay S, Feldman MD, Abels E, et al. Whole Slide Imaging Versus Microscopy for

Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). Am J Surg Pathol 2018;42:39–52.

[15] Babawale M, Gunavardhan A, Walker J, et al. Verification and Validation of Digital Pathology (Whole Slide Imaging) for Primary Histopathological Diagnosis: All Wales Experience. J Pathol Inform 2021;12:4.

[16] Kusta O, Rift CV, Risør T, Santoni-Rugiu E, Brodersen JB. Lost in digitization – A systematic review about the diagnostic test accuracy of digital pathology solutions. Journal of Pathology Informatics 2022;13:100136.

[17] Brodersen J, Schwartz LM, Heneghan C, O'Sullivan JW, Aronson JK, Woloshin S. Overdiagnosis: what it is and what it isn't. BMJ Evid Based Med 2018;23:1–3.

[18] Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington (DC): National Academies Press (US); 2015.

[19] Elmore JG, Longton GM, Pepe MS, et al. A Randomized Study Comparing Digital Imaging to Traditional Glass Slide Microscopy for Breast Biopsy and Cancer Diagnosis. J Pathol Inform 2017;8:12.

[20] Williams B, Hanby A, Millican-Slater R, et al. Digital pathology for primary diagnosis of screen-detected breast lesions - experimental data, validation and experience from four centres. Histopathology 2020;76:968–75.

[21] van Bergeijk SA, Stathonikos N, Ter Hoeve ND, et al. Deep learning supported mitoses counting on whole slide images: A pilot study for validating breast cancer grading in the clinical workflow. J Pathol Inform 2023;14:100316.

[22] Beckwith BA. Standards for Digital Pathology and Whole Slide Imaging. In: Kaplan KJ, Rao LKF, editors. Digital Pathology, Cham: Springer International Publishing; 2016, p. 87–97.

[23] Williams DKA, Graifman G, Hussain N, et al. Digital pathology, deep learning, and cancer: a narrative review. Transl Cancer Res 2024;13:2544–60.

[24] Song AH, Jaume G, Williamson DFK, et al. Artificial intelligence for digital and computational pathology. Nat Rev Bioeng 2023;1:930–49.

[25] Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. Diagn Pathol 2023;18:109.

[26] Zhu J, Liu M, Li X. Progress on deep learning in digital pathology of breast cancer: a narrative review. Gland Surg 2022;11:751–66.

[27] Rakha EA, El-Sayed ME, Lee AHS, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol 2008;26:3153–8.

[28] Curigliano G, Burstein HJ, Gnant M, et al. Understanding breast cancer complexity to improve patient outcomes: The St Gallen International Consensus Conference for the Primary Therapy of Individuals with Early Breast Cancer 2023. Ann Oncol 2023;34:970–86.

[29] Balic M, Thomssen C, Gnant M, Harbeck N. St. Gallen/Vienna 2023: Optimization of Treatment for Patients with Primary Breast Cancer - A Brief Summary of the Consensus Discussion. Breast Care (Basel) 2023;18:213–22.

[30] Chiang AC, Massagué J. Molecular basis of metastasis. N Engl J Med 2008;359:2814–23.

[31] Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. Science 2011;331:1559–64.

[32] Gupta GP, Massagué J. Cancer metastasis: building a framework. Cell 2006;127:679–95.

[33] Rakha EA, Miligy IM, Gorringe KL, et al. Invasion in breast lesions: the role of the epithelial-stroma barrier. Histopathology 2018;72:1075–83.

[34] Tan PH, Ellis I, Allison K, et al. The 2019 World Health Organization classification of tumours of the breast. Histopathology 2020;77:181–5.

[35] Rakha EA, Lee AHS, Evans AJ, et al. Tubular carcinoma of the breast: further evidence to support its excellent prognosis. J Clin Oncol 2010;28:99–104.

[36] Kim J, Kim JY, Lee H-B, et al. Characteristics and prognosis of 17 special histologic subtypes of invasive breast cancers according to World Health Organization classification: comparative analysis to invasive carcinoma of no special type. Breast Cancer Res Treat 2020;184:527–42.

[37] Giuliano AE, Edge SB, Hortobagyi GN. Eighth Edition of the AJCC Cancer Staging Manual: Breast Cancer. Ann Surg Oncol 2018;25:1783–5.

[38] Fitzgibbons PL, Page DL, Weaver D, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000;124:966–78.

[39] van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. Virchows Arch 2022;480:33–43.

[40] Wolff B. Histological grading in carcinoma of the breast. Br J Cancer 1966;20:36–40.

[41] Sistrunk WE, Maccarty WC. LIFE EXPECTANCY FOLLOWING RADICAL AMPUTATION FOR CARCINOMA OF THE BREAST: A CLINICAL AND PATHOLOGIC STUDY OF 218 CASES. Ann Surg 1922;75:61–9.

[42] Perry AC. The after-results of operations for malignant disease of the breast. Journal of British Surgery 1925;13:39–49.

[43] Greenough RB. Varying Degrees of Malignancy in Cancer of the Breast. The Journal of Cancer Research 1925;9:453–63.

[44] Patey DH, Scarff RW. THE POSITION OF HISTOLOGY IN THE PROGNOSIS OF CARCINOMA OF THE BREAST. The Lancet 1928;211:801–4.

[45] Haagensen CD. The Bases for the Histologic Grading of Carcinoma of the Breast. The American Journal of Cancer 1933;19:285–327.

[46] Bloom HJG. Prognosis in carcinoma of the breast. Br J Cancer 1950;4:259–88.

[47] Bloom HJG. Further studies on prognosis of breast carcinoma. Br J Cancer 1950;4:347–67.

[48] Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 1957;11:359–77.

[49] Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 1991;19:403–10.

[50] de Mascarel I, Bonichon F, Durand M, et al. Obvious peritumoral emboli: an elusive prognostic factor reappraised. Multivariate analysis of 1320 node-negative breast cancers. Eur J Cancer 1998;34:58–65.

[51] Kuhn E, Gambini D, Despini L, Asnaghi D, Runza L, Ferrero S. Updates on Lymphovascular Invasion in Breast Cancer. Biomedicines 2023;11:968.

[52] Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 2013;24:2206–23.

[53] Prat A, Cheang MCU, Martín M, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. J Clin Oncol 2013;31:203–9.

[54] Arpino G, Weiss H, Lee AV, et al. Estrogen receptor-positive, progesterone receptor-negative breast cancer: association with growth factor receptor expression and tamoxifen resistance. J Natl Cancer Inst 2005;97:1254–61.

[55] Rakha EA, Reis-Filho JS, Ellis IO. Basal-like breast cancer: a critical review. J Clin Oncol 2008;26:2568–81.

[56] Badve S, Dabbs DJ, Schnitt SJ, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Mod Pathol 2011;24:157–67.

[57] Nordenskjöld A, Fohlin H, Fornander T, Löfdahl B, Skoog L, Stål O. Progesterone receptor positivity is a predictor of long-term benefit from adjuvant tamoxifen treatment of estrogen receptor positive breast cancer. Breast Cancer Res Treat 2016;160:313–22.

[58] Maisonneuve P, Disalvatore D, Rotmensz N, et al. Proposed new clinicopathological surrogate definitions of luminal A and luminal B (HER2-negative) intrinsic breast cancer subtypes. Breast Cancer Res 2014;16:R65.

[59] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.

[60] Joensuu H, Isola J, Lundin M, et al. Amplification of erbB2 and erbB2 expression are superior to estrogen receptor status as risk factors for distant recurrence in pT1N0M0 breast cancer: a nationwide population-based study. Clin Cancer Res 2003;9:923–30.

[61] Slamon DJ, Godolphin W, Jones LA, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science 1989;244:707–12.

[62] Chia S, Norris B, Speers C, et al. Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers. J Clin Oncol 2008;26:5697–704.

[63] Franchet C, Djerroudi L, Maran-Gonzalez A, et al. [2021 update of the GEFPICS' recommendations for HER2 status assessment in invasive breast cancer in France]. Ann Pathol 2021;41:507–20.

[64] Bartlett JMS, Ellis IO, Dowsett M, et al. Human epidermal growth factor receptor 2 status correlates with lymph node involvement in patients with estrogen receptor (ER) negative, but with grade in those with ER-positive early-stage breast cancer suitable for cytotoxic chemotherapy. J Clin Oncol 2007;25:4423–30.

[65] Ignatiadis M, Sotiriou C. Breast cancer: Should we assess HER2 status by Oncotype DX? Nat Rev Clin Oncol 2011;9:12–4.

[66] Rodrigues MJ, Peron J, Frénel J-S, et al. Benefit of adjuvant trastuzumab-based chemotherapy in T1ab node-negative HER2-overexpressing breast carcinomas: a multicenter retrospective series. Ann Oncol 2013;24:916–24.

[67] Nielsen TO, Leung SCY, Rimm DL, et al. Assessment of Ki67 in Breast Cancer: Updated Recommendations From the International Ki67 in Breast Cancer Working Group. J Natl Cancer Inst 2021;113:808–19.

[68] Johnston SRD, Toi M, O'Shaughnessy J, et al. Abemaciclib plus endocrine therapy for hormone receptor-positive, HER2-negative, node-positive, high-risk early breast cancer (monarchE): results from a preplanned interim analysis of a randomised, open-label, phase 3 trial. Lancet Oncol 2023;24:77–90.

[69] Luporsi E, André F, Spyratos F, et al. Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review. Breast Cancer Res Treat 2012;132:895–915.

[70] de Azambuja E, Cardoso F, de Castro G, et al. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. Br J Cancer 2007;96:1504–13.

[71] Stuart-Harris R, Caldas C, Pinder SE, Pharoah P. Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. Breast 2008;17:323–34.

[72] Andre F, Ismaila N, Allison KH, et al. Biomarkers for Adjuvant Endocrine and Chemotherapy in Early-Stage Breast Cancer: ASCO Guideline Update. J Clin Oncol 2022;40:1816–37.

[73] Franchet C, Duprez-Paumier R, Lacroix-Triki M. [Molecular taxonomy of luminal breast cancer in 2015]. Bull Cancer 2015;102:S34-46.

[74] Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262–72.

[75] Metzger Filho O, Ignatiadis M, Sotiriou C. Genomic Grade Index: An important tool for assessing breast cancer tumor grade and prognosis. Crit Rev Oncol Hematol 2011;77:20–9.

[76] Spielmann M, Roché H, Delozier T, et al. Trastuzumab for patients with axillary-node-positive breast cancer: results of the FNCLCC-PACS 04 trial. J Clin Oncol 2009;27:6129–34.

[77] D'Hondt V, Canon J-L, Roca L, et al. UCBG 2-04: Long-term results of the PACS 04 trial evaluating adjuvant epirubicin plus docetaxel in node-positive breast cancer and trastuzumab in the human epidermal growth factor receptor 2-positive subgroup. Eur J Cancer 2019;122:91–100.

[78] Kerbrat P, Desmoulins I, Roca L, et al. Optimal duration of adjuvant chemotherapy for high-risk node-negative (N-) breast cancer patients: 6-year results of the prospective randomised multicentre phase III UNICANCER-PACS 05 trial (UCBG-0106). Eur J Cancer 2017;79:166–75.

[79] Campone M, Lacroix-Triki M, Roca L, et al. UCBG 2-08: 5-year efficacy results from the UNICANCER-PACS08 randomised phase III trial of adjuvant treatment with FEC100 and then either docetaxel or ixabepilone in patients with early-stage, poor prognosis breast cancer. Eur J Cancer 2018;103:184–94.

[80] Nakagawa K, Moukheiber L, Celi LA, et al. AI in Pathology: What could possibly go wrong? Semin Diagn Pathol 2023;40:100–8.

[81] Dehkharghanian T, Bidgoli AA, Riasatian A, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. Diagn Pathol 2023;18:67.

[82] Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. J Public Health Policy 2021;42:602–11.

[83] Van Giffen B, Herhausen D, Fahse T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research 2022;144:93–106.

[84] Vaidya A, Chen RJ, Williamson DFK, et al. Demographic bias in misdiagnosis by computational pathology models. Nat Med 2024;30:1174–90.

[85] Shi Y, Olsson LT, Hoadley KA, et al. Predicting early breast cancer recurrence from histopathological images in the Carolina Breast Cancer Study. NPJ Breast Cancer 2023;9:92.

[86] Dunn C, Brettle D, Cockroft M, Keating E, Revie C, Treanor D. Quantitative assessment of H&E staining for pathology: development and clinical evaluation of a novel system. Diagn Pathol 2024;19:42.

[87] Gray A, Wright A, Jackson P, Hale M, Treanor D. Quantification of histochemical stains using whole slide imaging: development of a method and demonstration of its usefulness in laboratory quality control. J Clin Pathol 2015;68:192–9.

[88] Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal

2019;58:101544.

[89] Hort M, Chen Z, Zhang JM, Harman M, Sarro F. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. ACM J Responsib Comput 2024;1:1–52.

[90] Asilian Bidgoli A, Rahnamayan S, Dehkharghanian T, Grami A, Tizhoosh HR. Bias reduction in representation of histopathology images using deep feature selection. Sci Rep 2022;12:19994.

[91] Degnan AJ, Ghobadi EH, Hardy P, et al. Perceptual and Interpretive Error in Diagnostic Radiology-Causes and Potential Solutions. Acad Radiol 2019;26:833–45.

[92] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4:e384–97.

[93] Kaczmarzyk JR, Saltz JH, Koo PK. Explainable AI for computational pathology identifies model limitations and tissue biomarkers. ArXiv 2024:arXiv:2409.03080v1.

[94] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell 2019;1:206–15.

[95] Evans H, Snead D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. NPJ Digit Med 2024;7:89.

[96] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med 2018;169:866–72.

[97] Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw Open 2023;6:e2335377.

[98] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176–82.

[99] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.

[100] Guitton T, Allaume P, Rabilloud N, et al. Artificial Intelligence in Predicting Microsatellite Instability and KRAS, BRAF Mutations from Whole-Slide Images in Colorectal Cancer: A Systematic Review. Diagnostics (Basel) 2023;14:99.

[101] Allaume P, Rabilloud N, Turlin B, et al. Artificial Intelligence-Based Opportunities in Liver Pathology-A Systematic Review. Diagnostics (Basel) 2023;13:1799.

[102] McGenity C, Clarke EL, Jennings C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med 2024;7:114.

[103] Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11:e048008.

[104] Yang B, Mallett S, Takwoingi Y, et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Ann Intern Med 2021;174:1592–9.

[105] Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. Ann Intern Med 2022;175:1010–8.

[106] Hercules SM, Alnajar M, Chen C, et al. Triple-negative breast cancer prevalence in Africa: a systematic review and meta-analysis. BMJ Open 2022;12:e055735.

[107] Hirko KA, Rocque G, Reasor E, et al. The impact of race and ethnicity in breast cancer-disparities and implications for precision oncology. BMC Med 2022;20:72.

[108] Wang Y, Song Y, Ma Z, Han X. Multidisciplinary considerations of fairness in medical AI: A scoping review. Int J Med Inform 2023;178:105175.

[109] Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023;7:719–42.

[110] Carey S, Pang A, Kamps M de. Fairness in AI for healthcare. Future Healthc J 2024;11:100177.

[111] Wang Y, Acs B, Robertson S, et al. Improved breast cancer histological grading using deep learning. Ann Oncol 2022;33:89–98.

[112] Border SP, Sarder P. From What to Why, the Growing Need for a Focus Shift Toward Explainability of AI in Digital Pathology. Front Physiol 2021;12:821217.

[113] Bera K, Katz I, Madabhushi A. Reimagining T Staging Through Artificial Intelligence and Machine Learning Image Processing Approaches in Digital Pathology. JCO Clin Cancer Inform 2020;4:1039–50.

[114] Amgad M, Hodge JM, Elsebaie MAT, et al. A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. Nat Med 2024;30:85–97.

[115] Diao JA, Wang JK, Chui WF, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat Commun 2021;12:1613.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Réduction du biais lié à la coloration des coupes histologiques de cancer du sein pour le développement d'outils de classification par intelligence artificielle

Mots clés : Cancer du sein, Intelligence artificielle, Réduction de biais, Apprentissage profond, Augmentation de données, Normalisation

Résumé : L'intelligence artificielle (IA) appliquée au diagnostic est en passe de bouleverser le métier de pathologiste. Les preuves de concept d'outils d'assistance au diagnostic abondent, et la commercialisation de tels outils a commencé. Cependant, un inconvénient fréquent des modèles d'IA en médecine, domaine dans lequel les jeux de données sont souvent relativement petits et dont l'annotation est soumise à caution, est leur tendance à prendre tout ou partie de leurs décisions sur la base de biais présents dans les jeux de données d'entraînement plutôt que sur des caractéristiques biologiques concrètes. Techniquement, il est bien connu qu'il existe une variabilité de coloration des images microscopiques du fait de la multiplicité des protocoles de coloration des tissus entre laboratoires, et dans le temps au sein d'un même laboratoire, constituant ainsi l'une des principales sources de biais dans l'apprentissage automatique en pathologie digitale. Pour remédier à cela, de nombreuses équipes ont publié des travaux sur la normalisation et l'augmentation des couleurs. Cependant, peu d'entre elles ont évalué leurs effets sur la réduction des biais et la généralisabilité des modèles. Dans cette étude, nous avons développé deux méthodes d'augmentation (AugmentHE) et de normalisation (HEnorm) des couleurs des images microscopiques en coloration standard (hématoxyline éosine (H&E)), et leurs effets sur la réduction des biais ont été monitorés. Ces deux méthodes ont également été comparées aux outils d'augmentation et de normalisation les plus largement utilisés dans la littérature.
Un jeu de données multicentrique conçu pour le grading histologique du cancer du sein a été utilisé. Grâce à cela, des modèles de classification ont été entraînés sur les images microscopiques de cancers du sein issues d'un centre unique avant d'évaluer leur performance sur des images provenant d'autres centres. Ce design a permis de monitorer de manière approfondie la réduction des biais tout en fournissant une évaluation précise des méthodes d'augmentation et de normalisation. AugmentHE entraînait une augmentation de 81 % de la dispersion des couleurs par rapport aux augmentations géométriques seules. De plus, tous les modèles de classification impliquant AugmentHE présentaient une augmentation significative de l'aire sous la courbe ROC (AUC) par rapport à la technique de RGB shift largement utilisée. Plus précisément, les modèles basés sur AugmentHE ont montré une augmentation d'au moins 0,14 de l'AUC par rapport aux modèles basés sur le RGB shift. En ce qui concerne la normalisation, HEnorm s'est révélé jusqu'à 78 fois plus rapide que les méthodes conventionnelles, tout en fournissant des résultats satisfaisants en termes de réduction des biais. En résumé, notre pipeline composé d'AugmentHE et de HEnorm a amélioré l'AUC sur des données biaisées jusqu'à 21,7 % par rapport aux augmentations usuelles. Les méthodes de normalisation conventionnelles couplées à AugmentHE ont donné des résultats similaires, bien qu'elles soient beaucoup plus lentes. En conclusion, nous avons validé un outil open source pouvant être utilisé dans tout projet de pathologie numérique basé sur l'apprentissage profond appliqué aux lames entières (WSI) colorées à l'H&E, qui réduit efficacement les biais liés à la coloration et pourrait, à terme, renforcer la confiance des pathologistes dans l'utilisation d'outils basés sur l'IA.

Title: Reduction of color-induced bias for artificial intelligence-based classification of breast cancer histological images

Key words: Breast cancer, Artificial intelligence, Bias mitigation, Deep learning, Data augmentation, Data normalization

Abstract: Artificial intelligence (AI)-assited diagnosis is an ongoing revolution in pathology. Many proofs of concept of diagnostic assistance tools are published, and some of them are already at the commercialization step. However, a frequent drawback of AI models in medicine, where datasets are often relatively small and annotations can be unreliable, is their propension to make part or all of their decisions rather on bias in training dataset than on concrete biological features. Technically, it is well known that variability in microscopic image staining arises from the multiplicity of tissue staining protocols across laboratories, and over time within a single laboratory, constituting one of the main sources of bias in machine learning for digital pathology. So as to deal with it, many teams have written about color normalization and augmentation methods. However, few of them have monitored their effects on bias reduction and model generalizability. In our study, two methods for stain augmentation (AugmentHE) and fast normalization (HEnorm) have been created and their effect on bias reduction has been monitored. Actually, they have also been compared to previously described strategies. To that end, a multicenter dataset created for breast cancer histological grading has been used. Thanks to it, classification models have been trained in a single center before assessing its performance in other centers images. This setting led to extensively monitor bias reduction while providing accurate insight of both augmentation and normalization methods. AugmentHE provided an 81% increase in color dispersion compared to geometric augmentations only. In addition, every classification model that involved AugmentHE presented a significant increase in the area under receiving operator characteristic curve (AUC) over the widely used RGB shift. More precisely, AugmentHE-based models showed at least 0.14 AUC increase over RGB shift-based models. Regarding normalization, HEnorm appeared to be up to 78x faster than conventional methods. It also provided satisfying results in terms of bias reduction. Altogether, our pipeline composed of AugmentHE and HEnorm improved AUC on biased data by up to 21.7% compared to usual augmentations. Conventional normalization methods coupled with AugmentHE yielded similar results while being much slower. In conclusion, we have validated an open-source tool that can be used in any deep learning-based digital pathology project on H&E whole slide images (WSI) that efficiently reduces stain-induced bias and later on might help increase pathologists' confidence when using AI-based products.