

PROPOSITION DE SUJET DE THESE, PROGRAMME INSTITUTS ET INITIATIVES - SCAI

TITRE DE LA THÈSE : DÉTECTION ET IDENTIFICATION DES ÉLÉMENTS D'INTERÊTS SUR DES IMAGES MICROSCOPIQUES DE FROTTIS SANGUINS.

DIRECTEUR DE THÈSE : XAVIER TANNIER

SIGLE ET NOM DU LABORATOIRE : LIMICS, UMRS 1142

NOM DE L'ÉQUIPE : LIMICS, UMRS 1142

ADRESSE : ESC. D, 2ÈME ÉTAGE, 15 RUE DE L'ÉCOLE DE MÉDECINE, 75006 PARIS

CO-ENCADRANTS : DANIEL STOCKHOLM

EQUIPE/LABORATOIRE DES CO-ENCADRANTS : UMRS 938 EQUIPE DÉVELOPPEMENT HÉMATOPOÏÉTIQUE ET LEUCÉMIQUE

Mots-clés : *Traitement d'images biomédicales, détection d'objets, Diagnostic automatique, Apprentissage profond, Réseaux de neurones à convolution, Hémopathies, Sang, Cellules, Microscopie*

Contexte

L'hémogramme est l'examen biologique le plus prescrit (60 millions par an en France). Cet examen clé du diagnostic des maladies du sang est réalisé par des automates d'hématimétrie qui mesurent par techniques physiques et chimiques le taux d'hémoglobine, le nombre de globules rouges, le nombre de plaquettes et le nombre de globules blancs. Sont également déterminés des paramètres spécifiques aux éléments figurés du sang, comme le volume des globules rouges et leur concentration en hémoglobine et le type de globules blancs qui composent la formule sanguine. Pour 10% des échantillons, les automates d'hématimétrie identifient des situations qui nécessitent de compléter l'examen. Le recours à un examen morphologique au microscope d'un frottis sanguin coloré est alors nécessaire. Cet examen morphologique demande une expertise de l'opérateur (technicien de laboratoire ou biologiste médical) tant dans la reconnaissance des éléments figurés que dans l'analyse globale du frottis sanguin et l'intégration de ces éléments à la situation clinico-biologique. L'analyse morphologique experte par l'opérateur s'effectue sur un nombre d'éléments donné, dépendant du type de diagnostic : l'examen de l'ensemble du frottis pour la détection d'agrégats de plaquettes et le diagnostic de pathologie érythrocytaire, l'examen de plusieurs milliers de globules rouges pour le diagnostic de paludisme, l'analyse de plusieurs dizaines de globules blancs pour le diagnostic d'une situation réactionnelle ou d'une hémopathie maligne. Des scanners de lames numérisant les images à divers grossissements permettent l'acquisition d'un nombre élevé soit de champs microscopiques, soit d'éléments figurés du sang isolés en quelques minutes. Il est donc possible de générer des banques d'images qui peuvent être annotées et servir à des systèmes d'apprentissage automatique à des fins de diagnostic de maladies du sang.

Objectif scientifique

Notre proposition consiste à mettre au point un système de reconnaissance automatique fondé sur des approches d'intelligence artificielle pour assister le diagnostic des maladies du sang. Le système utilisera des lames virtuelles obtenues à partir de frottis sanguins colorés en s'adressant aux trois catégories d'éléments figurés du sang : les plaquettes, les globules rouges et les globules blancs.

Justification de l'approche scientifique

Les travaux de la littérature proposent des outils de classification des différents globules blancs et d'identification de caractéristiques permettant de les sous-classifier en fonction de la pathologie. Par exemple, pour le diagnostic des leucémies aiguës, différentes stratégies de prétraitement et de segmentation des images permettent une extraction optimale des globules blancs d'intérêts, et certains modèles permettent d'identifier des caractéristiques corrélées avec la classification génomique de la leucémie [Eckardt, 2021]. Des travaux appliquant des réseaux de neurones convolutifs (CNN) à l'ensemble des éléments nucléés de frottis médullaires ont permis de développer des algorithmes capables d'identifier et classer les cellules du myélogramme avec robustesse [Matek, 2021]. Pour les globules blancs, un écueil est constitué par le fait qu'ils sont moins nombreux sur frottis de sang que sur frottis de moelle : un compromis doit être trouvé entre la rapidité, la quantité et la qualité d'images à acquérir et à analyser. Certaines équipes ont tiré parti de microscopes automatisés qui numérisent en quelques minutes un frottis sanguin. En utilisant l'analyseur DM96 CellaVision®, lui-même développé pour la classification par analyse d'image des éléments du sang, il a été possible d'entraîner un CNN pour distinguer entre 8 classes de globules blancs [Azevedo, 2019]. L'avènement de nouveaux scanners de lames permettant l'acquisition rapide d'images de résolution améliorée permet d'envisager d'alimenter des bases d'apprentissages de façon

beaucoup plus performante. Ceci a été réalisé par une équipe allemande [Haferlach, 2022] à partir de 10082 échantillons sanguins de routine numérisés sur des scanners MetaSystems® : 8425 images annotées par des experts cytologistes ont été utilisées pour entraîner un réseau de neurone [(Machine Learning model based on ImageNet pretrained Xception using Amazon Sagemaker (AS))] et permettre l'identification de 21 classes de cellules, comprenant la détection de blastes leucémiques ou de cellules lymphoïdes tumorales. N'ont pas été abordés par ce dernier travail les problèmes diagnostiques posés par les agrégats plaquettaires, la pathologie plaquettaire, et les pathologies impactant la morphologie des globules rouges.

Les principales difficultés rencontrées jusqu'ici pour automatiser la lecture des frottis sanguins sont donc les délais nécessaires pour acquérir et traiter des images microscopiques de résolution améliorée, la prise en compte de nombreux artefacts susceptibles d'induire des erreurs diagnostiques, la prise en compte de l'hétérogénéité de préparation des frottis sanguins (colorations variables selon le centre, le jour) et de méthodes d'acquisition (variable selon le type de scanner ou de microscope motorisé), et l'absence de développement de solutions répondant aux problèmes diagnostiques impliquant l'ensemble des éléments figurés du sang.

L'objectif est d'explorer plusieurs approches mêlant apprentissage automatique et traitement d'images pour la modélisation d'un système rapide et fiable, permettant un diagnostic automatique global sur des images microscopiques de frottis sanguin. La cible sera de répondre dans un délai inférieur à 30 minutes en pratique clinique quotidienne aux principales situations diagnostiques (>95%) nécessitant l'analyse cytologique sur frottis sanguin coloré des plaquettes, des globules rouges et/ou des globules blancs, en aval de l'hématimétrie.

Questions posées

- Peut-on utiliser des approches d'IA pour faciliter l'annotation des éléments figurés du sang dans un contexte d'exploration de multiples pathologies ?
- Comment les techniques d'apprentissage et de reconnaissance de formes permettent-elles de modéliser un système de diagnostic automatique pour les plaquettes, les globules rouges, les globules blancs ?
- Dans quelle mesure est-il possible, en se reposant sur des modèles d'IA, de distinguer, au sein d'un ensemble d'images microscopiques de frottis sanguins, des motifs correspondants aux principaux éléments du sang dans des conditions normales ou pathologiques ?
- Quels sont les avantages d'un tel système automatisé en routine clinique ?

Méthodes : L'étude sera réalisée à partir de données issues des laboratoires d'hématologie d'APHP-SU. Des frottis sanguins seront numérisés à l'aide d'un scanner Olympus VS200. Une collection de 80 lames numérisées est disponible et permettra d'amorcer le travail, sachant que chaque lame comporte des milliers d'éléments (plaquettes, globules rouges, globules blancs), chacun pouvant être considéré comme une image d'intérêt. De plus nous prévoyons de numériser un nombre de cas beaucoup plus important (plusieurs centaines voire milliers, le scanner permettant de numériser plusieurs dizaines de frottis par jour).

Dans un premier temps le doctorant participera à la détermination des stratégies d'annotation des images pour les plaquettes, les globules rouges et globules blancs des frottis sanguins. Le principe sera de déterminer avec les hématocytologistes les questions posées pour ces trois catégories d'éléments figurés du sang. Dans un second temps, les objectifs seront de parvenir à développer des outils permettant de détecter les agrégats de plaquettes, d'identifier des anomalies de taille, de forme et de contenu des globules rouges, de parvenir à classer l'ensemble des éléments nucléés et d'identifier les principales classes de globules blancs pathologiques que sont les blastes, les cellules lymphomateuses, les cellules lymphoïdes réactionnelles, les plasmocytes, les éléments de myélémie. Dans ce contexte une stratégie en deux étapes sera développée (séparément ou simultanément) :

- la détection des cellules d'intérêt en fonction du type de diagnostic (plaquettes/globules rouges/globules blancs)
- la classification des cellules

Nous nous appuyerons sur des résultats déjà obtenus par une thèse en cours (encadrement Renaud Piarroux et Xavier Tannier) dans le domaine de la détection de globules rouges infectés par *Plasmodium falciparum*, avec pour objectif de généraliser l'approche et les résultats, ainsi que l'évaluation en routine clinique.

Différentes stratégies pourront être appliquées, dans un premier temps, l'apprentissage par transfert et l'utilisation d'architectures neuronales pré-entraînées (ResNet, VGG) ayant déjà démontré de très bons scores de classification. Ces modèles ont prouvé leur intérêt pour des images microscopiques aux représentations complexes et très variables, malgré la différence importante de nature entre les jeux de données d'entraînement et les images ciblées.

Nous ajouterons les phénomènes de façon itérative, avec pour objectif l'amélioration continue et contrôlée des performances, en documentant précisément chaque étape et chaque évolution du jeu de données pour garantir la non-régression et permettre, pour chaque phénomène traité, d'évaluer l'impact possible en routine clinique. Dans ce contexte, les modèles pré-entraînés sont les mieux à même de garantir la robustesse des résultats, là où des modèles plus réduits et plus focalisés sur chaque problème pourraient être plus sensibles aux changements.

D'un point de vue scientifique, les questions que nous nous poserons seront justement liées à la généralisation des modèles à des phénomènes plus nombreux et plus variés :

- doit-on construire un même modèle de classification pouvant détecter et classifier tous les objets d'intérêt sur une lame, ou plutôt adopter une approche modulaire avec plusieurs systèmes ayant des rôles distincts ?
- comment contrôler les proportions de chaque classe d'intérêt dans le jeu d'entraînement, lorsque le nombre de classes augmente, avec des distributions et des niveaux de difficultés qui varient fortement ?
- comment améliorer encore le protocole de collecte de données, par exemple avec un système d'apprentissage actif collectant automatiquement les objets « difficiles » ?
- comment concilier la volonté de généraliser à des données toujours plus variées et plus proches de la réalité (différences de coloration, de qualité de l'image, de matériel), avec un besoin de résultats fiables et reproductibles pour une mise en œuvre efficace en routine clinique ?
- comment évaluer de façon pertinente un outil de diagnostic au niveau du patient, alors que le système lui-même s'applique au niveau d'une lame et même de chaque objet présent dans cette lame ?

Ces questions montrent le caractère indispensable d'une collaboration approfondie entre cliniciens et spécialistes de l'apprentissage statistiques. Dans ce contexte, les encadrants auront chacun un rôle clé tout en assurant la cohésion du projet : D Stockholm aura en charge l'encadrement des travaux en lien avec les annotations et les objectifs diagnostiques, X Tannier aura en charge l'encadrement du travail sur les systèmes d'apprentissage.

Adéquation à l'institut SCAI : Ce sujet s'inscrit pleinement dans l'axe « Santé » mis en avant par l'institut, par son utilisation des méthodes d'intelligence artificielle dans un contexte clinique. Si l'objectif est essentiellement celui d'une application médicale allant jusqu'à la mise en œuvre clinique, des questions méthodologiques restent posées.

Calendrier prévisionnel

Première année : état de l'art et mise en place d'un premier système de détection des éléments figurés du sang sur lames numérisées. Mise en place d'un système d'annotation intégrant si possible l'active learning. Application de diverses architectures neuronales pour la classification des éléments détectés. Evaluation de la qualité des approches.

Deuxième année : Augmentation de la variabilité des apprentissages et ajout de classes supplémentaires. Comparaison avec les méthodes de diagnostic standards . Rédaction d'une première publication.

Troisième année : Déploiement et premiers tests du système en routine clinique. Rédaction d'un deuxième article. Rédaction du mémoire de thèse.

Prérequis, formation : Connaissances (niveau master 2) en apprentissage automatique, reconnaissance de formes, analyse et traitement d'images. Intérêt fort pour les applications biomédicales. Connaissances en microscopie (optique)

Références:

Eckardt JN, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. *Leukemia*. 2022 Jan;36(1):111-118. doi:10.1038/s41375-021-01408-w

Matek C, et al. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood*. 2021 Nov 18;138(20):1917-1927. doi:10.1182/blood.2020010568

Haferlach T, et al. Automated Peripheral Blood Cell Differentiation Using Artificial Intelligence – a Study with More Than 10,000 Routine Samples in a Specialized Leukemia Laboratory <https://ash.confex.com/ash/2021/webprogram/Paper152447.html>, (ASH meeting, 2021)

Quelques publications des encadrants en lien avec le sujet :

- Anne-Cécile Normand, Aurélien Chaline, Noshine Mohammad, Alexandre Godmer, Aniss Acherar, Antoine Huguelin, Stéphane Ranque, Xavier Tannier, Renaud Piarroux. Identification of a clonal population of *Aspergillus flavus* by MALDI-TOF mass spectrometry using deep learning. *Scientific Reports*. Vol. 12, Issue 1575, January 2022.

- Cécile Nabet, Aniss Acherar, Antoine Huguenin, Xavier Tannier, Renaud Piarroux. Artificial Intelligence and Malaria. in *Artificial Intelligence in Medicine*, Lidströmer, Niklas and Ashrafian, Hutan (eds). Springer International Publishing, August 2021.

- Cécile Nabet, Aurélien Chaline, Jean-François Franetich, Jean-Yves Brossas, Noémie Shahmirian, Olivier Silvie, Xavier Tannier, Renaud Piarroux. Prediction of malaria transmission drivers in *Anopheles* mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry.

- Mehdi Hage-Sleiman, Ludovic Suner, Pierre Boncoeur, Michaël Chaquin, Fabrizia Favale, Jean-Alain Martignoles, Mara Memoli, Alexis Genthon,

Ollivier Legrand, Pierre Hirsch, François Delhommeau. Association entre lésions génomiques et signes de myélodysplasie dans les leucémies aiguës myéloïdes. Congrès annuel de la Société Française d'Hématologie, 2022

- Marinello M, Cosette J, Bogni C, Denard J, Buj-Bello A., Stockholm D. Characterization of Neuromuscular Junctions in Mice by Combined Confocal and Super-Resolution Microscopy. *J Vis Exp*. 2021 Dec 8;(178). doi: 10.3791/63032. PMID: 34958074.

- Aguanno D, Coquant G, Postal BG, Osinski C, Wieckowski M, Stockholm D, Grill JP, Carrière V, Seksik P, Thenet S. The intestinal quorum sensing 3-oxo-C12:2 Acyl homoserine lactone limits cytokine-induced tight junction disruption. *Tissue Barriers*. 2020 Oct 1;8(4):1832877. doi: 10.1080/21688370.2020.1832877. Epub 2020 Oct 26. PMID: 33100129; PMCID: PMC7714502.

- Cosette J, Moussy A, Paldi A, Stockholm D. Combination of imaging flow cytometry and time-lapse microscopy for the study of label-free morphology dynamics of hematopoietic cells. *Cytometry A*. 2017 Mar;91(3):254-260. doi: 10.1002/cyto.a.23064. Epub 2017 Mar 1.