Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale
Sciences de la **vie**
et de la **santé** | ED 414
Université de Strasbourg

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

Laboratoire ICube — UMR 7357

THÈSE présentée par :

Corentin MEYER

Soutenue le : **20 octobre 2023**

Pour obtenir le grade de : **Docteur de l'université de Strasbourg**
Discipline : **Sciences de la vie et de la santé**

Méthodes d'exploitation de données multimodales de patients par intelligence artificielle : application aux myopathies congénitales

THÈSE dirigée par :

M Olivier POCH	Directeur de recherche, CNRS
M Pierre COLLET	Professeur, Université de Strasbourg

RAPPORTEURS :

Mme Malika SMAIL-TABBONE	Maitre de Conférences, Université de Lorraine
M Antonio RAUSELL	Maitre de Conférences, Université de Paris

AUTRES MEMBRES DU JURY :

Mme Gisèle BONNE	Directrice de recherche, INSERM Paris
M Cédric WEMMERT	Professeur, Université de Strasbourg

Remerciements

Tout d'abord, je souhaite remercier les membres de mon jury de thèse qui me font l'honneur d'avoir accepté d'évaluer mes travaux de thèse. Merci au Dr Malika Smail-Tabbone, au Dr Antonio Rausell, au Dr Gisèle Bonne et au Pr Cédric Wemmert.

J'aimerais remercier l'ensemble des personnes qui ont rendus cette thèse possible, à commencer par l'équipe du CSTB, une équipe que je pourrais qualifier d'atypique, par sa diversité tant sur le plan scientifique que humain. Une équipe qui m'a permis de découvrir le monde de la recherche et de découvrir ce que je souhaitais vraiment faire. Alors merci aux permanents de l'équipe dans leur ensemble. Je souhaiterais remercier en particulier un premier bureau, celui de Laetitia, Arnaud et Nicolas. Pour toi Arnaud la lutte continue, merci pour ces discussions techniques, mais aussi politiques, culturelles et ces *memes* très obscurs. Nicolas je te souhaite soit d'être parmi les prochains, soit de réussir à ouvrir une salle d'escalade dans un autre pays.

J'aimerais avoir un mot particulier pour le duo Julie et Odile. Merci, Julie, de m'avoir donné l'opportunité de commencer en stage avec toi, puis de continuer en thèse dans cette équipe et enfin de m'avoir aidé dans cette dernière ligne droite. J'ai toujours admiré ta modestie et ta capacité à aller droit au but. Merci à toi Odile de m'avoir donné l'opportunité d'enseigner à l'ESBS c'était une expérience formidable qui m'a beaucoup aidé et je me suis même découvert une petite passion pour l'enseignement. En dehors du travail, vous êtes deux personnes que j'admire beaucoup sur le plan humain, avec vous, la bio-informatique à Strasbourg est sereine.

Je tiens à remercier aussi ceux que je nommerais les précaires, ces stagiaires et doctorants que j'ai pu rencontrer dans l' équipe et avec qui j'ai pu tisser du lien. Merci à Quentin, Tam'si, Nicolas S., Thomas, Romain et Arthur. Merci à Christelle, Hiba et Amani du "bureau d'à côté", vous êtes toutes les trois les prochaines à soutenir et je suis persuadé que cela va bien se passer. Enfin, merci aux personnes qui ont animé et habité le même bureau que moi, par le passé ou par le présent, merci à Célia, Lucille, Dorine et Alix, c'était un plaisir de partager ces moments avec vous qui ont donné lieu à des tranches de vie pleines de rebondissements, mais qui n'en resteront pas moins de bons souvenirs.

Merci à Jocelyn Laporte et son équipe ainsi qu'à l'équipe de Norma Romero pour ces collaborations qui ont porté leurs fruits à travers ces travaux de thèse. Vos connaissances extensives des myopathies ont été très stimulantes et nos échanges m'ont toujours permis de repartir l'esprit plein de nouvelles idées à mettre en place.

Enfin, je souhaite dédier un remerciement particulier pour mes directeurs de thèses et mes encadrants. Merci à Anne et Pierre pour leurs conseils et expertises en IA (ainsi qu'en philosophie !). Merci à ce duo de choc (c'est le cas de le dire), Olivier et Kirsley, dont l'énergie et l'animation

contrebalancent et n'ont d'égales que le calme et l'organisation d'Odile et Julie. Merci à Kirsley, toujours au four et au moulin entre les rédactions de *grant*, le développement et l'encadrement de stagiaires et doctorants, pour ses connaissances techniques et sa capacité à toujours aller plus loin dans les questionnements scientifiques. Merci à Olivier d'avoir été infatigablement derrière cette thèse du début à la fin, même dans les moments de flottement. Ainsi que pour cette capacité de raconter des histoires en continu dès que l'occasion se présente.

De manière générale, à tous ceux qui ont commencé comme collègues de travail, mais qui sont devenus bien plus que ça : merci, on sera amené à se revoir.

Avant d'être une aventure professionnelle, la thèse est avant tout une aventure personnelle. J'aimerais dédier cette partie à l'ensemble de mes proches qui n'ont pas été impliqués directement dans cette thèse, mais sans qui elle n'aurait jamais vu le jour.

À mes frères du groupe *Persepolis* et ADN : Adil, Arthur, Ernest, Keziah, Lucie, Malo et les autres. Merci. Du fond du cœur. Vous me rendez heureux et vous côtoyer tous les jours est un réel plaisir. J'espère que ça va perdurer le plus longtemps possible, on se sait. Le projet Japon est toujours dans un coin de ma tête.

À ceux qu'on côtoie moins à cause des trajets de vie, mais avec qui chaque retrouvaille est comme un retour à hier : merci, Bilal, Vincent, Baptiste et Victor. Vous avez vu, j'ai fini, je vais peut-être pouvoir sortir de la salle du temps.

Au Discord des doctorants français "*PhD Students*" et à sa communauté. C'est étrange de remercier une entité, mais j'aurais clairement arrêté la thèse sans ce refuge et sans avoir échangé (et râlé en cœur) avec tant de gens géniaux. Merci notamment à Anaïs et Floriane. Et merci à la délégation strasbourgeoise avec qui on a pu échanger tant de bons moments (et qui j'espère vont continuer). Merci au duo Alix et Alix, Emilien, Erin, Maï et Nato. Pour certains d'entre vous aussi, c'est bientôt votre tour. Un mot en particulier pour Alix n°2, camarade de conférences, de *gossip* et de discussions désastreuses, courage pour la suite, tu es bien entourée.

À tous ceux qui m'ont permis de sortir la tête de la thèse, souvent par l'escalade, mais pas que, merci. Merci à Louise, Magalie, Morgane, Pierre, Eloïse. Спасибо, Настя, надеюсь, что в Бретани нам удастся поесть морепродуктов !

Enfin, il me reste à remercier trois personnes spéciales, qui constituent ma famille. Je ne suis pas très doué pour le montrer, mais vous savez que je vous aime. Merci, Maman, merci, Olivier, merci, Coraline. Sans votre soutien inconditionnel toutes ces années durant, je ne serais évidemment pas ici aujourd'hui. Je sais que pour vous, le travail de thèse est assez obscur, j'avais peur de ne pas réussir à aller au bout, mais ne vous inquiétez pas après ça, c'est fini.

À tous mes proches que je remercie sincèrement à travers ces mots et qui savent que je ne suis pas forcément le plus habile pour ce genre d'exercice, si je devais résumer le fond de ma pensée en trois mots :

Vous êtes lumineux.

Sommaire

Résumé de Thèse

L'avènement des *Big Data* dans le domaine biomédical

Le terme "*Big Data*" fait référence à des ensembles de données extrêmement vastes, complexes et hétérogènes qui dépassent la capacité des outils de traitement de données traditionnels. Les *Big Data* sont caractérisées par les "5 V" : le volume, la variété, la vélocité, la véracité et la valeur. Les données biomédicales sont un exemple concret de *Big Data*. Elles sont multimodales : regroupant l'ensemble des données relatives à la santé humaine telles que les données génétiques, les dossiers cliniques, les images médicales ou les rapports d'analyses. Elles sont massives (ex. : taille des données de séquençage, imagerie à l'échelle du giga pixel, milliers de comptes rendus en texte libre…) et ont un flux important grâce à l'amélioration des techniques d'acquisition d'images, d'analyse et de séquençage. Ces données sont utilisées pour améliorer la compréhension des maladies et pour poser des diagnostics pour les patients. L'augmentation exponentielle des volumes, modalités et complexités des données biomédicales rend impossible leur traitement manuel et requiert donc l'utilisation d'outils adaptés pour traiter et extraire des informations pertinentes dans le cadre de la recherche médicale, comme les méthodes basées sur l'intelligence artificielle (IA).

Nouvelle génération d'IA pour le traitement du *Big Data* grâce aux réseaux de neurones

L'IA représente un ensemble de techniques permettant de créer des programmes simulant l'intelligence humaine. Une sous branche de l'IA nommée machine-learning (ML) regroupe des algorithmes permettant à un programme informatique d'accomplir une tâche en apprenant d'un jeu de données. Cependant, les techniques de ML traditionnelles ne peuvent apprendre que de données sous forme de tableaux, fermant la porte à l'exploitation de données plus complexes, comme les images ou les textes libres.

Cette dernière décennie, la popularisation des réseaux de neurones profonds (*Deep Neural Networks*, DNN), reposant sur le concept bio-inspiré de neurones, a permis l'exploitation de données complexes sans avoir besoin de connaissance à priori sur les données, c'est-à-dire sans devoir définir des descripteurs pertinents manuellement. Par exemple, grâce aux architectures de réseau neuronal convolutif (*Convolutional Neural Networks*, CNN), des modèles d'analyse d'images biomédicales ont été développés et mis à disposition de la communauté. Plus récemment encore, grâce aux architectures à base de modules d'attention (*Transformers*), des réseaux de neurones ont été entrainés pour comprendre du langage naturel (texte libre) et en extraire de l'information. L'année 2023 représente une année clé dans l'histoire du traitement de langage naturel (*Natural Language Processing*, NLP) avec le développement et la mise à disposition de modèles linguistiques de grande taille (*Large Language Models*, LLMs) généraux performants et accessibles tel que *GPT-3.5-turbo* (souvent nommé à tort *ChatGPT*) ou *LLAMA*. Ces méthodes permettent d'explorer de façon rétrospective et multimodale, l'ensemble des données biomédi-

cales acquises sur des patients sans avoir besoin de réaliser un travail manuel d'annotation trop important.

L'exemple des myopathies congénitales et la difficulté du diagnostic

Les myopathies congénitales (MC) sont une famille de maladies rares et génétiques. Cette maladie peut être causée par une mutation sur un panel de 35 gènes différents et présente une prévalence d'environ 1,5 pour 100 000, soit environ 1000 patients en France. Actuellement, les MC sont différenciées en cinq sous-types : myopathies à Némaline (NM), myopathies à *cores* (COM), myopathies centro-nucléaires (CNM), myopathies à disproportion congénitale des fibres (CFTD) et sans précision.

L'examen principal permettant la différenciation de ces sous-types de MC est l'histopathologie du muscle. Cet examen donne lieu à la rédaction d'un rapport d'analyse et permet de poser un diagnostic pour orienter le test génétique vers un groupe de gènes candidats. Cependant encore aujourd'hui, ce diagnostic est compliqué en raison de l'hétérogénéité des manifestations au niveau du muscle entre patients atteints d'un même sous-type de MC. Mais aussi en raison d'un chevauchement important des manifestations phénotypiques entre des sous-types de MC différents. La triple hétérogénéité des myopathies sur le plan clinique, histologique et génétique rendent difficile le diagnostic et l'orientation du test génétique. En raison de cette hétérogénéité et de la rareté de la maladie, 50 % des patients atteints de MC n'ont pas de diagnostic génétique à ce jour.

IMPatienT : un outil d'annotation et d'exploration de données multimodales de patients

Pendant ma thèse, en collaboration avec l'Institut de Myologie de Paris, j'ai développé une plateforme en ligne nommée **IMPatienT** (*Integrated digital Multimodal PATIENt daTa*) permettant de numériser et d'explorer les données de patients atteints de MC. Plusieurs centaines de rapports de biopsies musculaires ont été générés ces dernières décennies et cette masse de documents contient des informations expertisées sur les critères de différenciation des sous-types de MC. Cependant, ces documents sont sous la forme de texte libre semi-structuré. Nous avons donc utilisé une approche ontologique pour détecter et extraire les concepts clés dans ces rapports d'histologie et être capables d'en faire l'analyse statistique. La plateforme **IMPatienT** a été conçue en quatre modules reliés à une base de données unique.

Le premier module permet aux utilisateurs de créer leur propre vocabulaire standard (arborescence de termes ou de concepts) qui sera ensuite utilisé par le module 2. Au fur et à mesure que la base de données intègre des données patients, la définition des termes est enrichie avec les gènes et les diagnostics associés et les autres termes qui co-occurrent chez les patients. Le deuxième module permet de numériser des rapports d'histologie textuels et d'annoter automatiquement les concepts présents dans le rapport. Ces annotations peuvent être affinées par l'expert, ou en ajoutant des métadonnées liées aux rapports comme des symptômes cliniques, un gène muté, une variation et un diagnostic final. Enfin, un système d'aide à l'a décision est disponible afin suggérer un diagnostic sur la base de la similarité (méthodes bayésiennes) du profil du patient avec les patients déjà enregistrés dans la base de données d'**IMPatienT**. Le troisième module permet la segmentation automatique d'images histologiques, et leur annotation avec des termes issus du vocabulaire standard. Enfin, le dernier module correspond au tableau de bord de visualisation automatique. Il génère automatiquement des graphiques et tableaux permettant l'exploration statistique en temps réel des données enregistrées dans la base de données.

Développée avant la révolution des LLMs, **IMPatienT** est une plateforme permettant l'annotation semi-automatique et l'exploration de données multimodales de patients atteints de

MC. Les méthodes de détection des termes dans les rapports accélèrent le travail d'annotation, mais requièrent encore un travail manuel pour affiner les annotations réalisées. Ainsi, par la suite, j'ai exploré comment les LLMs peuvent automatiser ce travail d'annotation pour faciliter l'exploration des données.

Analyse de la base de données d'IMPatienT : classification des rapports par IA Explicable (xAI)

Le concept d'*eXplainable Artificial Intelligence* (xAI) se réfère à la capacité de comprendre et d'expliquer le fonctionnement des systèmes d'IA de manière claire et compréhensible pour les êtres humains. C'est une caractéristique importante des modèles IA notamment dans le domaine du diagnostic, car il est préférable pour l'Homme d'être en mesure de comprendre sur quels critères une prédiction est réalisée. Les 89 rapports annotés via **IMPatienT** ont été utilisés pour entrainer différents algorithmes d'IA dont les systèmes de classeurs (*Learning Classifier Systems*, LCS), considérés comme un système de référence en termes d'explicabilité. Nous avons comparé leurs performances sur des données réelles et avons obtenu une exactitude de classification de 83% pour la différenciation entre NM, COM et CNM. Nous avons aussi exploré de nouvelles façons de visualiser les connaissances contenues dans ces LCS pour faciliter l'extraction de connaissances.

NLMyo : Traitement de rapport textuel par modèles linguistiques de grande taille

Grâce aux développements récents de LLMs, nous avons pu développer **NLMyo** (*Natural Language Myopathies*), intégrant quatre outils pour permettre d'exploiter de manière totalement automatique et rapide un grand nombre de rapports textuels.

Le premier outil (*Anonymiser*) permet l'anonymisation des données, une étape essentielle pour travailler sur des données de santé. Grâce aux LLMs, nous pouvons détecter automatiquement les noms, prénoms et dates de naissance dans les documents et les censurer avant le traitement de ces données. Le second outil (*MyoExtract*) permet l'extraction automatique des métadonnées d'un rapport. Nous utilisons les LLMs pour extraire automatiquement le numéro de biopsie, l'âge du patient, le muscle prélevé et le diagnostic final. Nous extrayons ces informations dans un format standard (JSON) qui permet son traitement de manière automatique pour pré-remplir les champs des formulaires utilisés pour numériser les données de patients, par exemple dans **IMPatienT**. Avec le troisième outil (*MyoClassify*), nous avons exploré la possibilité de prédire un diagnostic de manière totalement automatique sans aucune annotation humaine à partir du texte brut du rapport et avons obtenu une justesse de classification de 65 % (versus 35 % pour le hasard). Cette classification est réalisée grâce à des techniques de vectorisation de phrases (*embedding*). L'*embedding* correspond à la transformation d'un texte en un unique vecteur numérique de grande taille (plusieurs centaines voire milliers de dimensions) capable de capturer le sens sémantique du texte. En appliquant cette méthode sur un corpus élargi de 192 rapports, nous avons pu entrainer un modèle d'IA capable de prédire le diagnostic associé à un rapport uniquement à partir de son *embedding*. Enfin, le quatrième outil (*MyoSearch*) exploite ces techniques d'*embedding* pour fournir un véritable moteur de recherche intelligent de patient. L'utilisateur peut formuler une requête en texte libre pour rechercher par exemple des patients ayant un symptôme spécifique. L'*embedding* de cette requête sera comparé à l'*embedding* de l'ensemble des phrases contenues dans les rapports histologiques, et les rapports avec les meilleurs scores seront présentés en premier. Cet outil permet de référencer et de rapidement trouver des patients ayant un diagnostic ou un profil symptomatique spécifique.

L'intégration future de ces méthodes dans **IMPatienT** facilitera la numérisation des données patients et permettra de gagner un temps important lors de l'annotation de ces patients dans la base de données.

Vers une génération de rapports automatique à partir d'imagerie avec MyoQuant

Grâce aux outils présentés précédemment, nous sommes en mesure d'exploiter les informations contenues dans les rapports histologiques de patients. Cependant, ces rapports sont rédigés à la main après observation de coupes de biopsie musculaire au microscope par un biologiste ou médecin. L'expertise manuelle des images de biopsie musculaire est couteuse en temps et elle n'est que qualitative : par exemple pour la centralisation nucléaire, un marqueur pathologique typique des MC, il sera noté qu'il y en a peu ou beaucoup, mais sans valeur numérique, car le comptage des fibres individuelles serait trop couteux en temps.

Nous avons développé **MyoQuant**, en collaboration avec l'équipe du Dr Jocelyn Laporte de l'IGBMC à Strasbourg, afin d'automatiser ce travail de comptage. Actuellement, **MyoQuant** peut quantifier des marqueurs pathologiques dans trois des cinq techniques de coloration réalisées en routine lors de la biopsie musculaire grâce à des systèmes d'IA. Pour la coloration Hématoxyline-Eosine (HE) qui met en évidence les noyaux cellulaires, l'algorithme est capable d'évaluer le niveau de centralisation de chaque noyau dans les fibres musculaires. Dans une fibre musculaire saine, les noyaux sont localisés en périphérie des fibres. En segmentant les fibres et les noyaux cellulaires d'une coupe histologique, nous calculons pour chaque noyau un score d'excentricité, représentatif de son niveau de centralisation, pour ensuite compter automatiquement le nombre de noyaux internalisés ou centralisés. Pour la coloration ATPase, qui colore de façon différentielle les fibres de type 1 et de type 2 et dont l'équilibre est modifié dans les MC, nous avons développé une méthode de ML capable de définir automatiquement un ou plusieurs seuils d'intensité qui permet de différencier et compter les fibres de chaque catégorie. Enfin, pour la coloration au Succinate Déshydrogénase (SDH) qui met en évidence l'activité oxydative des fibres, souvent anormale dans les COM, nous avons développé un réseau de neurones capable de détecter les fibres ayant une répartition mitochondriale anormale. Entrainé sur un total de 17 000 fibres musculaires issues de 17 souris modèles de MC, notre réseau de neurones obtient une exactitude de classification de 93 %.

Nous souhaitons à présent étendre le champ de détection de **MyoQuant** en développant des méthodes pour détecter les agrégats protéiques des colorations au Trichrome de Gomori (TG) ainsi que les *cores* dans la coloration NADH, pour permettre à terme de générer automatiquement un rapport de biopsie musculaire plus précis.

Conclusions et Perspectives

Dans le cadre de cette thèse, j'ai eu l'occasion de développer plusieurs outils permettant d'exploiter des données multimodales de patients par approches IA. Avec **IMPatienT**, j'ai créé une plateforme d'annotation et d'exploration de rapports histologiques de patients qui a permis de numériser une centaine de rapports de patients. La base de données d'**IMPatienT** a été utilisée pour évaluer les performances de plusieurs approches xIA pour la prédiction des MC. Ensuite, avec **NLMyo**, j'ai exploré comment les récentes avancées en NLP grâce aux LLMs pouvaient faciliter et accélérer l'annotation et la classification de ces rapports textuels. Enfin, avec **MyoQuant**, j'ai développé un outil qui permet d'accélérer le processus d'évaluation des images de biopsies musculaires avec comme but, à terme, d'être capable de générer un rapport histologique de façon automatique.

Les avancées en IA ouvrent la voie pour une amélioration du diagnostic des maladies rares. Les outils que j'ai développés sont un exemple de science translationnelle. Sur le plan diagnostic, ces outils peuvent permettre un gain de temps pour les praticiens via l'automatisation de l'extraction d'information lors de l'évaluation des résultats d'analyse. Sur le plan de la recherche, ils peuvent permettre la découverte de nouveaux critères de diagnostic potentiellement importants. Il serait maintenant intéressant de compiler l'ensemble de ces outils dans une plateforme unique, cohérente et clé en main utilisable par la communauté pour démocratiser l'usage de ces outils. De plus, l'intégration de méthodes d'analyse de données génomique complèterait la palette d'outils mis à disposition.

Publications et communications

Papier soumis dans Journal of Neuromuscular Diseases : Meyer, C., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. (2022). IMPatienT : An Integrated web application to digitize, process and explore Multimodal PATIENt daTa. (p. 2022.04.08.487635). bioRxiv. https ://doi.org/10.1101/2022.04.08.487635.

Poster :

IMPatienT : an integrated web application to digitize, process and explore multimodal patient data – Meyer, C., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. - ED Days 2022 - Strasbourg France

Methods for the exploitation of multimodal patient data using artificial intelligence : application to congenital myopathies - Meyer, C., Giraud Q., Vernay B., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. – ISMB/ECCB 2023 - Lyon France

Communication orale (demo) et Poster : IMPatienT : an integrated web application to digitize, process and explore multimodal patient data - Meyer, C., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. – JOBIM 2022 - Rennes France

Communication orale :

MyoQuant : a tool to automatically quantify pathological features in muscle fiber histology images – Meyer, C., Giraud Q., Vernay B., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. - JSFM 2022 - Toulouse France

MyoQuant : a tool to automatically quantify pathological features in muscle fiber histology images - Meyer, C., Giraud Q., Vernay B., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. - Journée Fondation Maladie Rare 2022 - Strasbourg France

MyoQuant : a tool to automatically quantify pathological features in muscle fiber histology images - Meyer, C., Giraud Q., Vernay B., Romero, N., Evangelista, T., Cadot, B., Laporte, J., Jeannin-Girardon, A., Collet, P., Chennen, K., & Poch, O. - ED Days 2023 - Strasbourg France

Table des figures

Liste des tableaux

Glossaire

A | B | C | D | F | G | H | I | K | L | M | N | O | R | S | T | W | X

A

API *Application Programming Interface*. 40, 70, 126, 129, 135, 166
ATP Adénosine triphosphate. 45

B

BOQA Bayesian Ontology Query Algorithm : algorithme de prédiction basé sur des informations de fréquence de termes issue d'ontologies. 110

C

CERN Conseil européen pour la recherche nucléaire. 73
CFTD Myopathies à disproportion congénitale des fibres. xii, 47, 51
CNM Myopathies centro-nucléaires. xii, xiii, 47, 49, 110, 112, 114, 126, 129, 132, 135, 138, 149
CNN Réseau neuronal convolutif. xi, xvii, 24, 26, 27, 29, 34, 150
COM Myopathies à *cores*. xii–xiv, 47–49, 51, 110, 112, 114, 126, 129, 132

D

DNN Réseaux de neurones profonds. xi, xvii, xxi, 17–24, 29, 33, 150
DOI *Digital Object Identifier*, DOI. 73
DSE Dossiers de Santé Électroniques. 5, 8

F

FDA Food and Drug Administration. 17

G

GPU *Graphical Processing Unit* (carte graphique). 22, 23, 69, 141, 152, 159, 165

H

HE Hématoxyline-Eosine. xiv, 53, 61, 68, 138, 141, 159
HGVS *Human Genome Variation Society*. 62, 77, 106
HPO *Human Phenotype Ontology*. 61, 77, 106

I

IA Intelligence artificielle. xi, xiii–xv, xxvii, xxviii, 3, 8, 12, 15, 17–19, 34, 35, 37, 41, 54, 55, 59, 72, 106, 107, 137, 138, 141, 149, 150, 153, 159, 163, 165–168
IGBMC Institut de génétique et de biologie moléculaire et cellulaire. 59, 140, 149
IMPatienT *Integrated digital Multimodal PATIENt daTa*. xii–xiv, xxvii, 55, 61, 62, 64, 68, 71, 77, 78, 106, 107, 109, 118, 119, 126, 132, 137, 163–167

K

KDE *Kernel density estimation*. 145

L

LCS Systèmes de classeurs. xiii, 15, 64, 109, 112, 113, 115, 118
LLMs Modèles linguistiques de grande taille. xi–xiv, 23, 34, 35, 37, 40, 55, 69, 70, 72, 77, 106, 119, 121, 123–126, 129, 134, 135, 165, 166

M

MC Myopathies congénitales. xii–xiv, 46–49, 51, 52, 54, 55, 106, 109, 114, 115, 126, 137, 138, 140, 144, 149, 150, 159, 164
MCC coefficient de corrélation de Matthew *(Matthew Correlation Coefficient, MCC)*. 67, 112
ML Machine-Learning. xi, xiv, xxvii, 3, 8–10, 12–16, 40, 55, 62, 64, 77, 109, 110, 112
MyoQuant outil de quantificaiton automatique de caractéristiques pathologiques dans les images histologiques des fibres musculaires. xiv, xxviii, 55, 68, 69, 71, 72, 137, 138, 140, 141, 159, 163, 165–167

N

NGS Séquençage de nouvelle généraiton (*Next Generation Sequencing* (NGS)). 52
NHS Service de santé national du Royaume-Uni (*National Health Service, NHS*. 7
NLMyo *Natural Language Myopathies*. xiii, xiv, xxvii, 55, 62, 69–72, 119, 134, 137, 154, 163, 165–167
NLP Traitement de langage naturel. xi, xiv, 34, 35, 119
NM Myopathies à némaline. xii, xiii, 47–49, 110, 112, 114, 115, 126, 129, 132, 134
NMD atteintes neuromusculaires (*neuromuscular diseases*, NMD). 45–47, 52, 53

O

OCR Optical Character Recognition. 69, 126, 133
ORDO *Orphanet Rare Disease Ontology*. 61, 77, 106

R

RegEx Expression régulières. 121, 123, 124
RGPD Règlement général sur la protection des données. 119, 166, 167
RLHF apprentissage par renforcement avec retour humain (*Reinforcement Learning from Human Feedback, RLHF*). 35

S

SDH Succinate déshydrogénase. xiv, 53, 61, 138, 149, 150, 152, 154, 159
SSL apprentissage auto-supervisé (*self-supervised learning, SSL*). 35, 37

SVM *Support Vector Machine*. 112

T

TG Trichrome de Gomori. xiv, 53, 159

W

WSI *Whole Slide Images*. 141, 145, 149, 152

X

xIA Intelligence Artificielle Explicable. xiii, xiv, 12, 165

Avant-Propos

L'objectif de cette thèse est de développer des méthodes et des outils d'intelligence artificielle adaptés à l'exploitation des données biomédicales multimodales, en particulier les comptes rendus médicaux et les données d'imagerie avec une application aux myopathies congénitales. Cette thèse a donné lieu au développement de trois outils mettant à disposition les moyens nécessaires à l'exploitation de ces données.

L'**introduction** est séparée en trois chapitres (**chapitres 1-3**). Les deux premiers chapitres posent le contexte informatique et biologique dans lequel s'inscrivent ces travaux.

Le **chapitre 1** s'intéresse à l'émergence du *Big-Data* dans le cadre des données biomédicales, une partie de la diversité des données biomédicales et décrit les principes de bases et les approches traditionnelles utilisées en intelligence artificielle pour analyser les données.

Le **chapitre 2** s'intéresse plus spécifiquement aux intelligence artificielle (IA) de type réseaux de neurones et comment cette nouvelle technologie transforme la manière de traiter les données biomédicales multimodales non structurées avec des exemples d'IA déjà existantes en imagerie et en analyse de texte.

Enfin, le **chapitre 3** pose le contexte biologique de la thèse avec une présentation de notre cas d'application : les myopathies congénitales. On y retrouve d'abord une présentation générale du muscle puis une description des myopathies congénitales, de leur diagnostic et des données générées suite à celui-ci.

Le **chapitre 4 matériels et méthodes** décrit l'ensemble des ressources utilisées pour développer ces outils. On y retrouvera d'abord une description des ressources biologiques comme les ontologies et la source de données biomédicales utilisées. Puis l'ensemble des ressources informatiques nécessaires au développement des outils est détaillé, tel que les algorithmes utilisés, les bibliothèques de programmation, le matériel informatique, les modèles pré-existants ainsi que les méthodes d'évaluation des performances. Enfin, dans une dernière section, l'accent est mis sur les outils permettant de rendre ces travaux de recherche *open-source* et reproductibles.

La partie **contributions** est divisée en quatre chapitres (**chapitres 5-8**).

Le **chapitre 5** présente **IMPatienT** (*Integrated digital Multimodal PATIENt daTa*), le premier outil développé durant cette thèse, qui est une application web d'annotation de comptes rendus et d'images de patients. Cet outil a fait l'objet d'un article soumis dans un journal à comité de lecture dont le manuscrit est intégré au chapitre.

Le **chapitre 6** s'intéresse à l'analyse détaillée de 89 comptes rendus de patients intégrés dans **IMPatienT** et leur classification par modèles de machine-learning (ML) classiques et ML explicable.

Le **chapitre 7** présente **NLMyo** (*Natural Language Myopathies*), une boite à outil basé sur les modèles linguistiques de grandes tailles pour l'exploitation automatique des comptes rendus

médicaux. Cette boite à outils permet d'anonymiser et d'extraire de l'information de comptes rendus textuels, ainsi que de les classer et de créer un moteur de recherche de symptômes automatiquement.

Enfin, le **chapitre 8** présente la possibilité de générer de manière automatique des comptes rendus de biopsie grâce à des méthodes de quantification par IA. Pour cela, nous avons développé **MyoQuant**, un outil de quantification automatique de marqueurs pathologiques sur des biopsies musculaires.

Pour finir le **chapitre, 9 discussions et ouvertures** traitent des principaux challenges, limites et perspectives des outils développés. Tout d'abord, des perspectives biologiques sont traitées avec l'intégration des données génomiques et la mise en relation des différentes modalités. Des perspectives techniques sont également considérées, notamment concernant l'explicabilité de l'IA, du déploiement de ce type d'approche à grande échelle en termes de ressources informatiques et des questions de législation en termes de données de santé et de traitement automatique. Une ouverture supplémentaire est faite sur la possibilité de création d'un produit regroupant l'ensemble de ces outils dans le cadre du concours *Mature Your PhD* organisé par la Satt Connectus.

Première partie

INTRODUCTION

Chapitre 1

Big Data, données biomédicales et apprentissage automatique

> *"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."*
>
> – Atul Butte, 2012

La numérisation du monde a permis la production et l'accumulation de données de façon exponentielle. Dans le domaine de la santé, les avancées technologiques comme les technologies de séquençage, d'imagerie ou les dossiers médicaux électroniques ont permis au cours du temps la capture de données précieuses pouvant améliorer la compréhension des maladies et la prise en charge des patients.

Dans le même temps, les technologies d'analyse de données massives (*Big Data*) se sont développées notamment grâce à l'apprentissage automatique (ML), une branche de l'IA permettant à des algorithmes informatiques d'apprendre à partir des données. Cette massification des données biomédicales et le développement des technologies d'analyse de données permettent d'entrevoir un monde où les soins de santé pourraient être personnalisés, préventifs, prédictifs et participatifs, c'est la médecine 4P.

Dans ce premier chapitre d'introduction, nous présenterons d'abord les *Big Data*, la variété des données biomédicales et comment ces données sont utiles aux des patients et à l'ensemble des acteurs de la médecine. Puis, dans une seconde partie, nous présenterons les concepts principaux du ML pour le traitement de ces données.

1.1 Les données biomédicales : des *Big Data* au service des patients

Dans cette section, nous présentons trois exemples de données biomédicales qui rendent compte de la diversité de données biomédicales de patients et donc du défi de leur exploitation. Nous apportons une attention particulière aux données d'imagerie, aux comptes rendus médicaux et aux données génétiques, car ces données sont essentielles au diagnostic des myopathies congénitales. Cependant, cette présentation ne se veut pas exhaustive, les données biomédicales sont un domaine très vaste qui regroupe de nombreux autres types de données.

1.1.1 Définition du *Big Data*

Le terme *Big Data* est utilisé pour faire référence à l'immense quantité de données complexes et hétérogènes produites par cette révolution numérique et le développement des technologies haut débit (De Mauro et al., 2016). La définition des *Big Data* a été enrichie au fur et à mesure des années, d'abord par 3 concepts puis 5 et même jusqu'à 7 (García et Álvarez-Fernández, 2022). La définition la plus commune aujourd'hui des *Big Data* se compose des 5 "V" : le volume, la variété, la vélocité, la véracité et la valeur (Ishwarappa et Anuradha, 2015). Ainsi pour être considérée comme *Big Data*, les données doivent : (i) être volumineuses, du gigaoctet à l'exaoctet (ii) être variées, c'est-à-dire multimodales (iii) avoir une vélocité de création et de traitement importante (iv) être véraces, c'est-à-dire valide et (v) avoir une forte valeur ajoutée, c'est-à-dire qu'elles doivent être utiles. (García et Álvarez-Fernández, 2022).

Les données biomédicales sont en adéquation avec cette définition (Zheng et al., 2021). Grâce aux améliorations en techniques d'acquisitions (séquençage de nouvelle génération, imagerie haute-résolution) elles sont volumineuses et possèdent une forte vélocité. Ces données sont variées, contenant des informations sur les plans génétique, phénotypique et histologique par exemple. De plus, elles ont une certaine véracité couplée à une forte valeur ajoutée. En effet, ce sont de manière générale des données générées par des experts (médecin ou biologiste) et liées à des patients (et donc fortement valorisées). Il est alors juste de qualifier les données biomédicales de *Big Data* (Sonawane et al., 2019). Dans la prochaine section, nous allons voir en détail les différentes modalités de données biomédicales de patients, leurs acquisitions et les challenges que présente leur analyse.

1.1.2 Variété des données biomédicales

Les données biomédicales sont par nature multimodales (Acosta et al., 2022), le diagnostic d'un patient se réalise par l'intégration de différents niveaux d'informations. Tout d'abord, il y a les données phénotypiques, listant les symptômes et autres caractéristiques du patient après certains examens médicaux. Ces données peuvent être sous la forme de texte libre, rédigé par le praticien de santé. Dans une gamme de nouvelles technologies et de nouvelles possibilités, il y a les données d'imagerie, issues d'examens complémentaires pour mieux caractériser l'atteinte du patient (échographie, IRM, histopathologie). Enfin, les données génétiques et omiques sont nécessaires, notamment dans le cadre de maladies génétiques, mais aussi en cancérologie, pour cibler les dysfonctionnements d'origine génétique (données de types séquences). Ces données sous forme de texte libre, d'imagerie et de séquences impliquent des techniques d'acquisition, de traitement et des difficultés propres.

1.1.2.1 Données textuelles et comptes rendus médicaux

Si les examens médicaux classiques donnent en général lieu à des données numérisées et structurées (tels que les bilans sanguins et les électrocardiogrammes), le recours à des examens spécialisés pour compléter les informations sur le patient peut donner lieu à la rédaction de comptes rendus médicaux en texte libre contenant des informations expertisées, denses et hautement valorisées. Les données en texte libre sont très communes dans le cadre des données de santé. L'accumulation au cours du temps d'archives de comptes rendus médicaux riches d'informations à forte valeur ajoutée est une source d'informations importante qui reste à explorer.

Dans le cadre de cette volonté d'explorer ces données les dossiers de santé électroniques (DSE) (*electronic health records*, *EHR*) sont des outils pour numériser ces comptes rendus textuels afin de les centraliser et les exploiter (Graber et al., 2017). Cependant, le développement d'outils pour numériser et exploiter les comptes rendus en texte libre reste difficile. La compréhension du texte libre par un programme informatique est une tâche ardue. C'est pourquoi la majorité des solutions de DSE demandent une phase d'annotation manuelle (remplissage de formulaires et de champs) par l'utilisateur pour numériser les données, ce qui en pratique est rarement réalisé en raison de la difficulté technique (temps d'annotation, utilisation de jargon et acronymes spécifiques).

1.1.2.2 Données d'imagerie

Le développement de l'imagerie a permis une diversification des techniques (IRM, radiologie, échographie, microscopie optique et électronique, imagerie 2D et 3D...) tout en améliorant leur résolution et précision de capture et en réduisant les couts associés (Abdallah, 2017; Prakash et al., 2022; Sheppard, 2021). Ainsi l'imagerie médicale est devenue un examen de routine pour le diagnostic de diverses pathologies. Cette production de données d'imagerie en routine et de grande résolution a donné lieu à une massification des données d'imagerie. Néanmoins, en histologie et microscopie à haute résolution, il est difficile pour un clinicien d'évaluer manuellement ces données de manière exhaustive, d'explorer l'ensemble des informations disponibles. Le développement d'outils capable d'analyser et de quantifier les éléments d'intérêt sur les données d'imagerie est donc un enjeu majeur (Tchito Tchapga et al., 2021) pour à la fois accélérer l'évaluation des données, mais aussi pour améliorer la précision des cliniciens. Par exemple, dans le cadre de la microscopie photonique, il est maintenant courant d'utiliser des scanneurs de lames complètes, générant ainsi des images à l'échelle du gigaoctet par lame. L'analyse de ces images est extrêmement couteuse en temps si l'on veut réaliser une évaluation manuelle exhaustive et un comptage des caractéristiques pathologiques en vue d'un diagnostic.

1.1.2.3 Données génétiques et omiques

Enfin, les progrès en terme de technologies de séquençage, grâce notamment aux technologies de seconde génération à lectures courtes (technologie Illumina) et aux technologies de troisième génération à lectures longues (technologie PacBio et Nanopore), ont permis l'accès à l'ensemble des informations génétiques de l'Homme. De plus, la baisse des couts de séquençage rend possible l'utilisation de séquençage de génome complet pour le diagnostic de maladies génétiques chez les patients (Rabbani et al., 2012).

Plus récemment, les techniques dites "omiques" (figure 1.1, Momeni et al., 2020) sont utilisées pour mieux comprendre les pathologiques telles que les technologies de transcriptomiques (expression ARN des gènes), épigénétiques, protéomiques (expression protéique des gènes) et métabolomiques (étude des métabolites). Ces technologies permettent d'obtenir une vue globale

des mécanismes biologiques qui opèrent au sein d'un tissu. Elles permettent d'accéder à tous les niveaux d'information chez un individu malade et sont devenues omniprésentes notamment en cancérologie, car elles ouvrent la voie à la médecine de précision (R. Chen et Snyder, 2013; Dai et Shen, 2022; Raufaste-Cazavieille et al., 2022).

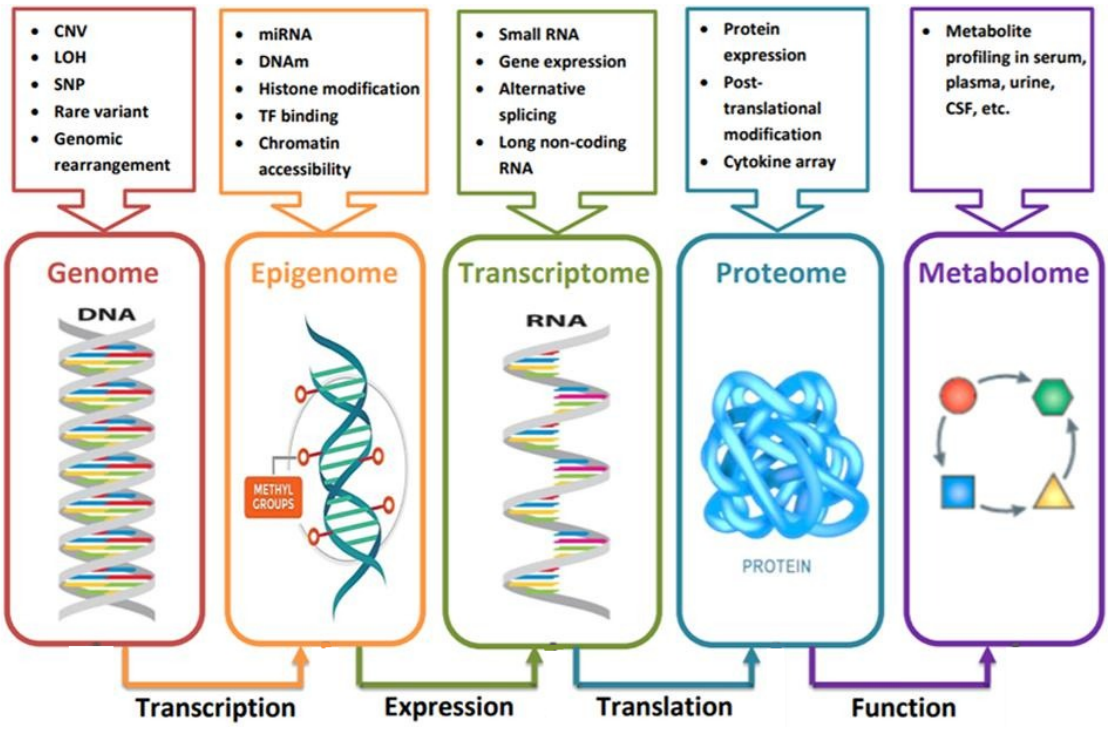


Figure 1.1 – **Schéma des différentes méthodes de séquençages "omiques"**. Les méthodes omiques donnent accès à une vue globale des mécanismes biologiques dans les tissus biologiques. (Modifié de Momeni et al., 2020)

Les données de séquençage sont massives, le génome humain mesurant environ 3,1 milliards de paires de bases (bp), et le séquençage d'un génome unique avec une profondeur de 50X (nécessaire pour la détection de mutation génétique) représente un minimum de 150 milliards de paires de bases lues et stockées, pour un individu. Ces données de séquence massives requièrent des outils spécifiques et un matériel informatique adapté à leur traitement. Outre l'aspect massif de ces données, la détection de mutations pathogènes est complexe. L'identification du gène responsable d'une maladie génétique reste un challenge lors du diagnostic, même avec les données de séquençage complètes.

1.1.3 Ressources et initiatives en données biomédicales et maladies rares

Plusieurs ressources et initiatives ont été créées dans un effort de structuration des données biomédicales à travers la communauté scientifique. Une méthode de structuration de données très utilisée repose sur le concept d'ontologie. Les ontologies sont une représentation structurée et formelle d'entités et de leurs relations. Ces ontologies sont en quelque sorte un vocabulaire standard et universel développé de manière collaborative pour former un consensus sur les termes et leurs définitions. Parmi les ontologies les plus utilisées dans le domaine biomédical on retrouve

Human Phenotype Ontology, (HPO, Robinson et al., 2008, Köhler et al., 2021), qui défini la liste des termes utilisés pour décrire les symptômes cliniques ou encore la *Gene Ontology* (GO, The Gene Ontology Consortium et al., 2023) qui défini la fonction des gènes, les processus biologies et les localisations cellulaires. À ce jour, le *NCBI BioPortal* (https://bioportal.bioontology.org/) référence 1062 ontologies biomédicales développées et utilisées. Les ontologies utilisées dans le cadre de cette thèse sont présentées plus en détail dans le chapitre" Matériels et Méthodes" (4.2). Ces ontologies sont très utiles pour structurer les données biomédicales grâce à des méthodes d'annotation qui vont reconnaitre les termes issus de ces ontologies dans les données non structurées telles que les textes libres.

En plus des ontologies, plusieurs bases de connaissances ont été développées, notamment dans la recherche sur les maladies rares. Parmi elles, on retrouve la base de données *Online Mendelian Inheritance in Man (OMIM* Amberger et al., 2015) qui répertorie toutes les maladies héréditaires humaines connues, le gène associé et les phénotypes observés. Il y a également *RD-Connect* (Laurie et al., 2022), un projet européen permettant de connecter les bases de données existantes, les registres de patients et les biobanques en une plateforme centrale et disponible. Enfin, Orphanet (Maiella et al., 2013) est une initiative française qui développe à la fois une ontologie des maladies rares (*ORDO*) mais qui est aussi un immense portail d'information destiné aux maladies rares qui répertorie des données sur les professionnels de santé, les centres experts, les études cliniques, de médicaments orphelins et de projets de recherches. L'ensemble de ces ressources représentent une base de connaissance importante qui peut être utile à la structuration et à l'exploitation automatique des données biomédicales.

1.1.4 Collecte et utilisation des données biomédicales au service du patient

Le Royaume-Uni est un pays pionnier dans la collecte et la mise à disposition de façon massive de données biomédicales à travers le service de santé national (*National Health Service, NHS*). Cela s'illustre par exemple par le projet "100 000 génomes" lancé en 2012 qui a pour but de séquencer 100 000 génomes de patients Anglais pour améliorer la recherche et le diagnostic de maladies rares, certains cancers et maladies infectieuses (Nunn et al., 2019). En avril 2022 a été publié le rapport de 112 pages intitulé : "*Better, broader, safer : using health data for research and analysis*" (Ben Goldacre, 2022), écrit par le professeur Ben Goldacre missionné par le NHS. Ce rapport met en évidence les challenges et la stratégie à adopter pour une collecte et un usage à grande échelle de données biomédicales de patients. Le projet OpenSAFELY (https://www.opensafely.org/), fondé par Ben Goldacre, est un exemple concret d'utilisation de données biomédicales au service de la recherche et de la prise en charge de patients. Ce projet, créé en juin 2020 pour lutter contre la pandémie de COVID-19, met à disposition des chercheurs des outils et des données biomédicales massives de patients. À ce jour, ce projet a permis la publication de plus de 80 publications scientifiques de recherches réalisées à partir de ces données. Des initiatives similaires mettent en évidence l'utilité des *Big Data* biomédicales comme catalyseur de découvertes scientifiques, tel que le projet "Big Data to Knowledge" fondé par le *National Institutes of Health (NIH)* (Toga et al., 2015). Cette collecte et utilisation des données biomédicales se généralise et d'autres initiatives sont à l'œuvre notamment aux Etats-Unis avec le projet TopMed (https://topmed.nhlbi.nih.gov/) par exemple.

Outre la phase de collecte, la difficulté dans l'exploitation des données biomédicales réside dans la disponibilité de techniques d'analyse adaptées (Ismail et al., 2020; Wang et al., 2019). Les données biomédicales étant volumineuses, complexes et multimodales, leurs explorations manuelles ou *via* des ontologies et techniques de statistiques classiques ne sont pas suffisantes. Une solution réside dans l'utilisation de l'intelligence artificielle, et plus spécifiquement de

la branche nommée l'apprentissage automatique (*machine-learning*, ML) pour construire des systèmes capables d'exploiter ces données. De nombreuses techniques d'analyse des DSE reposent aujourd'hui sur l'utilisation de modèles IA (de Mello et al., 2022; Li et al., 2022; X. Yang et al., 2022).

1.2 Apprentissage automatique pour le traitement des données biomédicales

Le ML est une branche de l'IA qui regroupe un ensemble d'algorithmes capables d'accomplir une tâche en apprenant d'un jeu de données. Dans cette section, nous allons définir les concepts de base du ML tels que le format des données, les tâches qui peuvent être accomplies, les méthodes d'apprentissage et les principaux algorithmes utilisés.

1.2.1 Les formats et partitionnements des données

Les données sont le point critique des techniques de ML. Elles représentent l'ensemble des informations utilisées par un algorithme de ML pour réaliser son apprentissage et réaliser des prédictions. Pour être utilisables par les algorithmes de ML, les données doivent être structurées. Le tableau 1.1 présente un exemple de structure d'un jeu de données exploitable par un algorithme de ML. Les données sont sous la forme de tableau où chaque ligne représente une observation (par exemple un patient) et chaque colonne représente un descripteur (nommé *feature* en anglais, par exemple le rythme cardiaque, la présence d'une toux chez le patient, la présence d'antécédents de diabète dans la famille...). Enfin, la dernière colonne représente le label, que l'on souhaite prédire dans le cadre de l'entrainement de notre modèle.

ID Patient	Rythme Cardiaque (bpm)	Toux	Diabète	Diagnostic
1	86	non	non	Sain
2	65	?	non	Sain
3	59	non	?	Sain
4	95	oui	non	Malade
5	101	oui	oui	Malade

Tableau 1.1 – **Exemple de tableau de données fictives de patients**. Les "?" indiquent l'absence d'information.

Cette contrainte sur la structure nécessaire du jeu de données pour les algorithmes de ML met en évidence les limites de leurs utilisations pour l'analyse de données non structurées telles que le texte libre, les données d'imagerie. Il est nécessaire en amont de structurer ces données à travers des descripteurs pertinents pour les exploiter.

De plus, il est nécessaire de partitionner ce jeu de données sous forme de deux tableaux : les données d'apprentissage et les données de test. Les données d'apprentissage sont les données qui vont être utilisées par l'algorithme de ML pour réaliser son entrainement, c'est-à-dire pour apprendre à réaliser la tâche définie (prédiction du label par exemple). Le jeu de test quant à lui contient des données qui n'ont jamais été présentées au modèle au cours de l'apprentissage. Le modèle entrainé va alors prédire le label du jeu de test et les prédictions réalisées sont comparées aux labels réels. Cela permet d'évaluer les performances d'un entrainement. Pour donner un ordre de grandeur, il est commun d'utiliser 80% des données comme jeu d'entrainement et 20% des données restantes comme jeu de test.

Pour finir, il existe un troisième partitionnement des données optionnel nommé jeu de validation. Le jeu de validation est réalisé en général en prenant 10% des données d'entrainement. Ce jeu de validation permet d'évaluer le modèle au cours de l'entrainement et ajuster ses paramètres. Ceci permet de s'assurer que l'entrainement progresse correctement avant de tester les performances à la fin de l'entrainement sur le jeu de test.

1.2.2 Les différentes tâches que le *machine-learning* peut accomplir

Les algorithmes de ML peuvent accomplir de multiples tâches, dont quatre principales : (i) les algorithmes de classification (ii) de régression (iii) de *clustering* et (iv) de réduction de dimensionnalité.

Les tâches de classification sont les plus communes. Il s'agit ici d'apprendre à prédire une classe ou un label pour un point de donnée. Par exemple, dans le cadre de données biomédicales, il peut s'agir de la prédiction d'un diagnostic parmi une liste de maladies. Cette classification peut-être binaire (2 classes uniquement, par exemple sain vs malade) ou multi-classe (plus de 2 classes, par exemple faire la différence entre 10 maladies potentielles). Enfin, cette classification peut aussi être multilabel, c'est-à-dire que l'on peut prédire plusieurs classes pour un point de donnée. Par exemple, si l'on construit un algorithme capable de prédire la fonction d'un gène, il est utile d'avoir un système de classification multilabel pour prédire les différentes fonctions dans lesquels un seul et même gène est impliqué. Parmi les algorithmes de ML capables de faire de la classification, on retrouve de nombreux outils (décrit dans la section 1.2.4) tels que les méthodes bayésiennes, les méthodes à base d'arbres (arbre de décision, forêt aléatoire) ou encore les systèmes de classeurs.

Les tâches de régression ne cherchent pas à prédire une catégorie, mais une valeur numérique. Par exemple, on peut construire un modèle ML capable de prédire le prix d'une maison ou encore la pression sanguine d'un patient, dans ces cas-là on cherche à prédire une valeur numérique continue. Les algorithmes à base d'arbres (arbre de décision, forêt aléatoire) sont aussi capables de réaliser des tâches de régression et on retrouve aussi d'autres algorithmes tels que la régression Lasso, Ridge et régression linéaire, qui est l'algorithme de base pour les tâches de régression.

Les tâches de *clustering* cherchent à regrouper les points de données similaires en sous-groupes (c'est-à-dire en *clusters*). Les techniques de *clustering* sont utilisées dans le domaine biomédical pour analyser les données d'expression génétique par exemple. À partir de l'expression des gènes d'une cohorte de patients, il est possible d'utiliser des algorithmes de *clustering* pour stratifier des sous-groupes de patients ayant un profil d'expression similaire (en cancérologie par exemple). Les algorithmes classiques de *clustering* sont l'algorithme K-means (MacQueen, 1967), DBSCAN (Ester et al., 1996) et le *clustering* hiérarchique (Cohen-Addad et al., 2017).

Pour finir, les tâches de réduction de dimensionnalité consistent à réduire le nombre de variables aléatoires d'un jeu de données en obtenant un ensemble de variables principales. Typiquement, les données à haute dimensionnalité comme les données transcriptomiques (expression de plusieurs dizaines de milliers d'ARN) sont complexes à analyser et présentent des problèmes spécifiques à cette haute dimensionnalité, connus sous le nom de la malédiction de la dimension (*curse of dimensionality*). Les techniques de réduction de dimensionnalité tendent à atténuer ce problème. Les algorithmes de réduction de dimensionnalité sont typiquement utilisés après une étape de *clustering* pour observer graphiquement les *clusters* obtenus en un graphique 2D. Pour reprendre l'exemple précédent, après une analyse transcriptomique, une étape de réduction de dimensionnalité permet de visualiser le principal axe de différenciation des échantillons. La réduction de dimensionnalité peut aussi être utilisée pour sélectionner les descripteurs les plus pertinents pour une tâche de classification. Les algorithmes de réduction de dimensionnalité communément utilisés sont la PCA (Maćkiewicz et Ratajczak, 1993), le t-SNE (Maaten et

Hinton, 2008) et UMAP (McInnes et al., 2020).

1.2.3 Apprentissage supervisé, non supervisé et par renforcement

Les différentes tâches présentées peuvent se regrouper sous trois méthodes d'apprentissages différentes : l'apprentissage supervisé, non supervisé et par renforcement.

Les tâches de classification et de régression sont possibles grâce à l'apprentissage supervisé. En apprentissage supervisé, le modèle est entrainé sur des données labellisées, c'est-à-dire des données pour lesquels on connait déjà le résultat attendu (diagnostic par exemple). Ainsi le modèle est entrainé à reproduire ces labels automatiquement. Les tâches de *clustering* et de réduction de dimensionnalité sont possibles grâce à l'apprentissage non supervisé. En apprentissage non supervisé, les labels des données ne sont pas connus par l'algorithme. L'objectif est donc de découvrir la structure cachée des données à partir des descripteurs. Ainsi le modèle essaie de déterminer des sous-groupes ou des regroupements de dimensions qu'il détermine comme pertinents, mais sans connaitre le résultat réel attendu.

Enfin, l'apprentissage par renforcement est moins connu et représente une méthode d'apprentissage où un agent (modèle) apprend à se comporter dans un environnement donné, recevant des pénalités et des récompenses en fonction de ses actions. Typiquement, un modèle apprenant à jouer à un jeu d'échecs représente une tâche d'apprentissage par renforcement. Dans un cadre biomédical, un système d'apprentissage par renforcement peut être utile par exemple pour un système d'examen médical intelligent qui va proposer des symptômes à vérifier chez le patient en fonction des observations déjà enregistrées. Ainsi, on a un environnement (les observations réalisées chez le patient) et des actions à réaliser par le modèle de ML (proposer des symptômes ou examens à vérifier). Les algorithmes utilisés pour ce type d'apprentissage peuvent être des réseaux de neurones, les méthodes de *Q-learning* ou encore les systèmes de classeurs.

La figure 1.2 représente schématiquement la classification des tâches, des modes d'apprentissages et des différents cas d'applications et algorithmes associés.

1.2.4 Algorithmes de *machine-learning* et explicabilité

Enfin, dans cette section de présentation des outils de ML, nous allons présenter la notion d'explicabilité des algorithmes de ML et son importance dans le domaine biomédicale. L'utilisation d'algorithmes explicables est cruciale dans le domaine de la santé, afin de pouvoir les utiliser en conditions réelles. Pour cela, les modèles de ML entrainés et utilisés doivent être transparents, c'est-à-dire qu'on doit être capable de pouvoir comprendre et évaluer leurs prédictions. Cette transparence permet une confiance accrue dans le modèle à la fois par le patient et par le personnel médical, mais aussi d'éviter de potentielles erreurs. En effet lors d'un désaccord entre le personnel médical et une prédiction, il est alors possible dans le cadre d'un modèle transparent d'évaluer les raisons du désaccord pour prendre la meilleure décision possible pour le patient. Ainsi nous allons voir quelques exemples d'algorithmes de ML couramment utilisés pour voir leur fonctionnement et leur niveau d'explicabilité.

1.2.4.1 Le concept d'explicabilité

D'après l'essai philosophique "*Studies in the logic of explanation*" de Carl G. Hempel et Paul Oppenheim en 1948 (Hempel et Oppenheim, 1948), le concept d'explication scientifique peut se résumer en une équation :

$$\sum C + \sum L = E$$

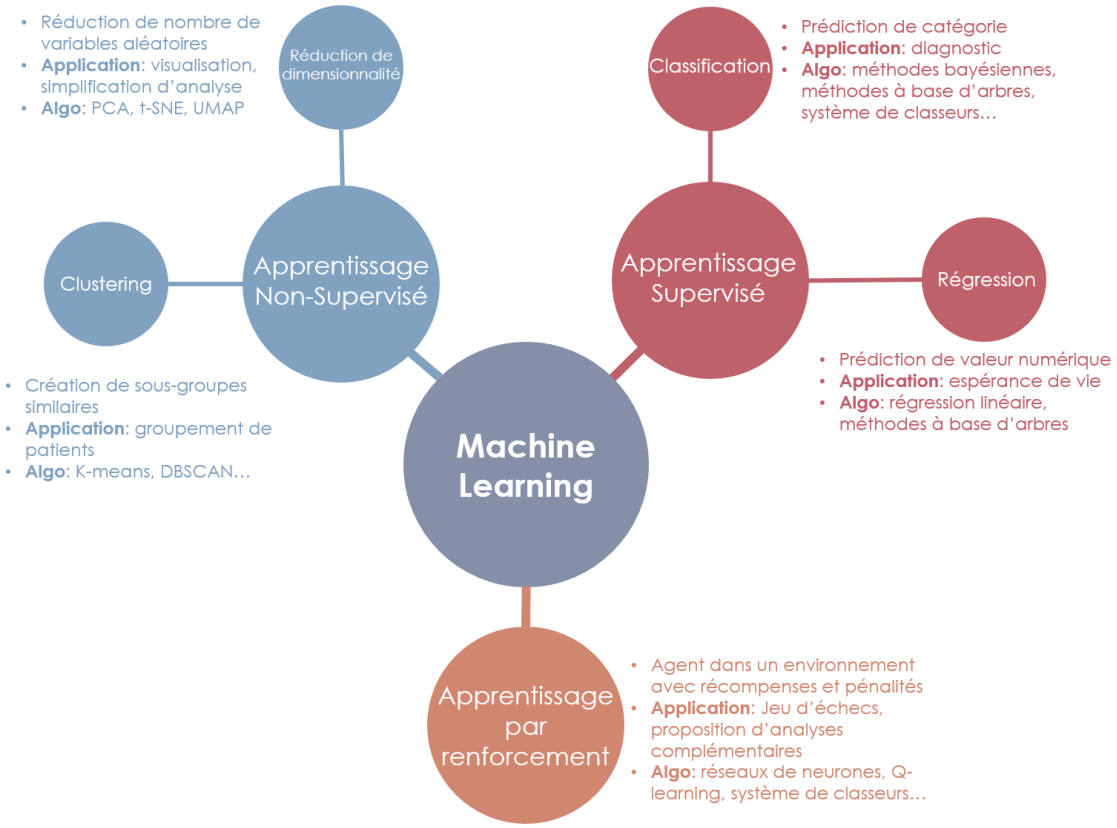


FIGURE 1.2 – **Schéma de la classification des tâches, des modes d'apprentissage et des différents cas d'applications et algorithmes associés en *machine-learning*.** Il y a trois grands types d'apprentissage en machine-learning : l'apprentissage supervisé, non supervisé et par renforcement. Le diagnostic de patient par machine-learning est une tâche de classification par apprentissage supervisé.

Dans cette équation, C représente l'ensemble des conditions antérieures et L représente l'ensemble des lois générales. La somme des conditions et des lois permet de produire E, l'évènement ou le phénomène observé. Les termes de gauche représentent ce qu'on nomme l'*explanans* (l'expliquant) et le terme de droite est référé comme l'*explanandum* (l'explicable). L'équation mathématique implique que l'on peut la lire dans les deux sens. C'est-à-dire qu'en connaissant E (le phénomène), nous pouvons déduire C et L (les conditions et lois) et nous réalisons donc une explication scientifique. À l'inverse, en connaissant C et L (les conditions et lois), nous pouvons déduire E (le phénomène) et nous faisons alors une prédiction. Optimalement, une explication est adéquate si l'*explanans* permet de prédire totalement le phénomène observé.

Appliqué au *machine-learning*, C représente alors les points de données et leurs descripteurs (conditions initiales), L représente notre modèle de ML et ses règles internes, tandis que E représente la prédiction du modèle.

Les récentes recherches en IA ont amené à l'émergence du domaine d'*eXplainable Artificial Intelligence* (xAI). L'xIA cherche à concevoir des méthodes pour rendre les modèles d'IA et de ML plus transparents et explicables, ce qui est critique dans le cadre de l'application de ces modèles dans des domaines à haut risque, comme le domaine médical (Arrieta et al., 2019). L'objectif est donc de concevoir des méthodes de ML dont on est capable de comprendre et d'évaluer les prédictions de manière intelligible.

La figure 1.3 présente le concept de compromis entre les performances des algorithmes et leur niveau d'explicabilité. De manière générale, plus un algorithme est performant d'un point de vue prédictif, moins il est explicable. Dès lors, le domaine de l'xIA va chercher : (i) à améliorer l'explicabilité des algorithmes performants et peu explicables ou (ii) à améliorer les performances des modèles les plus explicables.

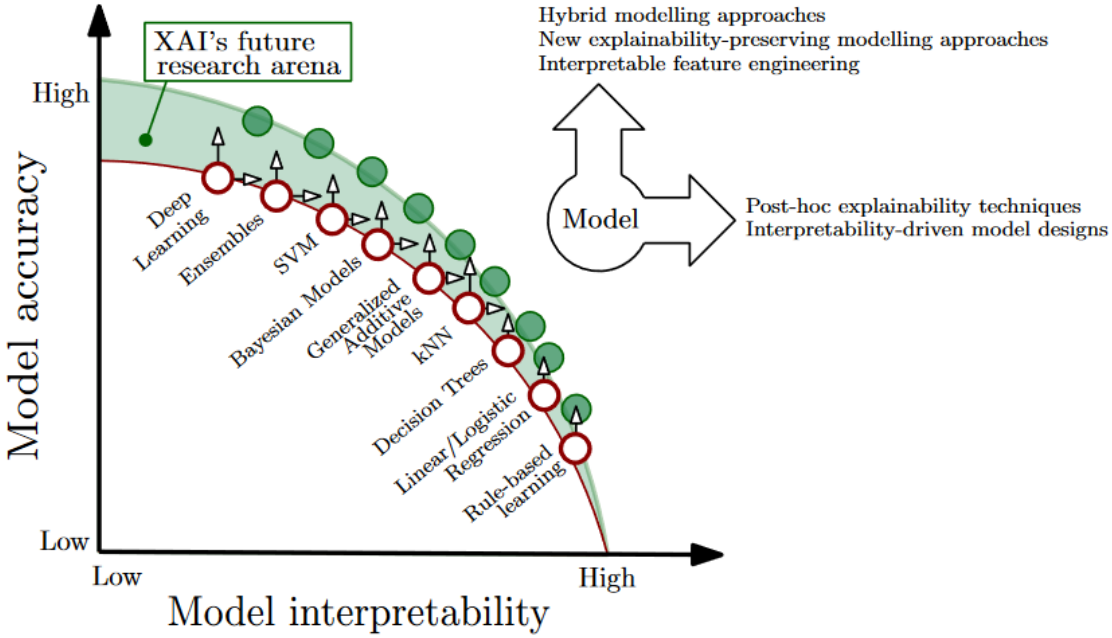


Figure 1.3 – **Compromis entre interprétabilité du modèle et performance d'algorithmes de ML.** Représentation de la zone où réside le potentiel d'amélioration des techniques et outils d'IA explicable (xAI) (Arrieta et al., 2019)

Au regard de l'explicabilité d'un modèle en ML, il y a deux catégories d'algorithmes : (i) les algorithmes transparents par design, c'est-à-dire directement explicables et (ii) les algorithmes sous forme de boites noires, dont l'explicabilité n'est accessible que grâce à des méthodes *post-hoc*. Dans les prochaines sous-sections, nous allons présenter le fonctionnement des algorithmes les plus utilisés de façon non exhaustif pour chaque catégorie. Nous allons ainsi voir les méthodes bayésiennes, les arbres de décisions et les systèmes de classeurs comme méthodes transparentes. Puis les méthodes de forêts aléatoires et de *boosting* seront présentées comme méthodes à explicabilité *post-hoc*.

1.2.4.2 Méthodes bayésiennes

Les méthodes bayésiennes naïves reposent sur le théorème de Bayes sur les probabilités conditionnelles avec une hypothèse d'indépendance forte (c'est-à-dire naïve) entre les descripteurs. Pour chaque descripteur, la probabilité d'une classe est calculée en fonction de la valeur du descripteur. Grâce à cette caractéristique probabiliste, le modèle est capable de fournir une prédiction et une mesure de l'incertitude associée. Les méthodes bayésiennes sont transparentes et donc explicables, car il est possible de décomposer la contribution de chaque descripteur lors d'une prédiction.

1.2.4.3 Méthodes à base d'arbres

Les arbres de décision sont des modèles classiques en ML. Cette méthode cherche à produire un arbre (un graphe dirigé acyclique) où chaque nœud représente une condition basée sur les descripteurs des données. En fonction de la valeur de ces descripteurs, l'arbre suit le chemin correspondant vers un sous-nœud, finalement aboutissant à une prédiction au niveau d'un nœud feuille. Cette méthode est intrinsèquement explicable, car il est possible de représenter graphiquement l'arbre et de suivre le processus de prédiction pour un point de données en appliquant les conditions comme des règles logiques. Pour illustrer ce concept, la figure 1.4 présente un exemple simplifié d'arbre de décision comportant trois nœuds, chaque nœud représentant une condition basée sur un descripteur, et deux classes possibles comme prédictions au niveau des nœuds feuilles.

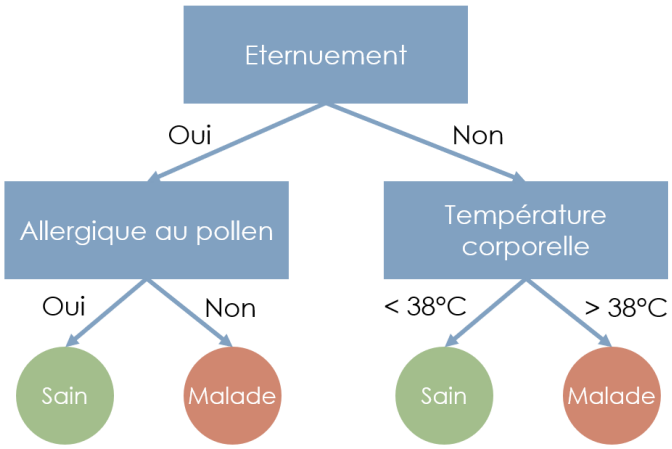


FIGURE 1.4 – **Exemple de schéma d'arbre de décision.** Cet arbre est composé de 3 questions (3 nœuds pour 3 descripteurs, représentés par des rectangles bleus) et 2 classes (disques verts et orange).

Une évolution des arbres de décision pour les rendre plus complexes et performants est nommée la méthode de la forêt aléatoire qui est une méthode ensembliste des arbres de décision. La figure 1.5 présente le fonctionnement de la forêt aléatoire. Cela consiste à construire plusieurs arbres de décision sur des sous-ensembles des données, puis de combiner les prédictions de ces arbres. Chaque arbre de décision ainsi généré va établir une prédiction et la prédiction finale correspondra à la majorité des votes de chaque arbre. Les méthodes de *boosting* (XGBoost (T. Chen et Guestrin, 2016), LGBoost (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2019)) sont similaires aux forêts aléatoires à la différence que chaque arbre n'est pas construit indépendamment, mais il cherche à corriger les erreurs du précédent, de manière séquentielle.

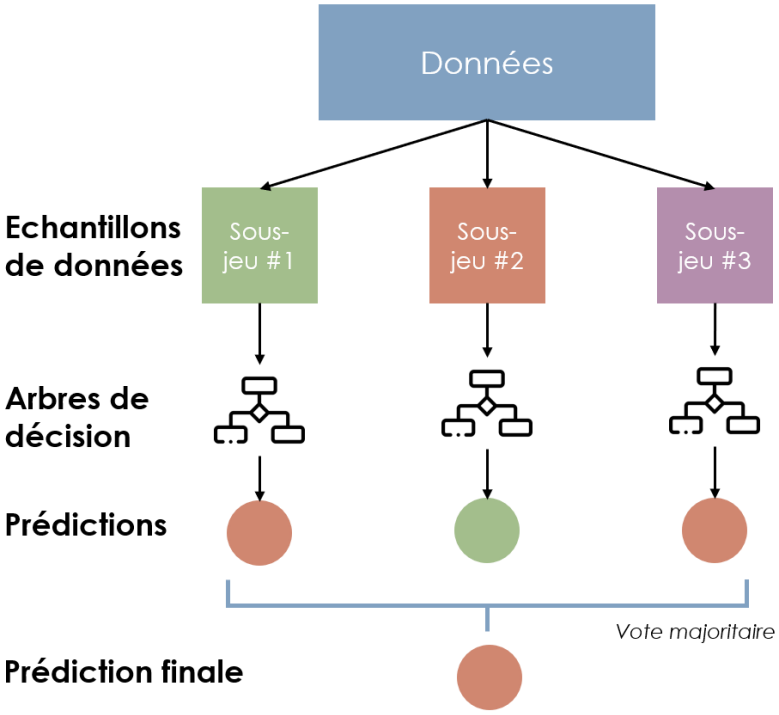


Figure 1.5 – **Schéma de fonctionnement de l'algorithme de forêt aléatoire**. Le jeu de données est divisé en plusieurs sous-jeux de données dans lesquels un sous-ensemble de descripteurs est conservé aléatoirement. Un arbre de décision est entrainé sur chacun de ces sous-jeux de données. La prédiction finale correspond à la majorité des prédictions de chaque arbre de décision.

Les forêts aléatoires sont très utilisées en ML et spécifiquement en biologie. En exemple d'application dans le domaine biomédicale, l'outil MISTIC (*MISsense deleTeriousness predICtor*, Chennen et al., 2020) développé dans notre équipe utilise des méthodes de forêts aléatoires pour calculer la pathogénicité des variants génétiques. Cependant, la forêt aléatoire et les méthodes de *boosting* perdent en explicabilité. En effet, en raison du grand nombre d'arbres de décision générés, il n'est plus possible de suivre aisément le processus de prédiction suivant une suite de règles logiques pour un point de donnée. Il faut alors utiliser des méthodes d'explicabilité *post-hoc* qui consistent à calculer l'importance numérique de chaque descripteur pour une prédiction donnée.

1.2.4.4 Systèmes de classeurs (LCS)

Les systèmes de classeurs (*Learning Classifier Systems*, LCS) sont des algorithmes de ML parmi les plus transparents et explicables (*"rules-based learning"* figure 1.3, Arrieta et al., 2019). Ces systèmes fonctionnent sur le principe d'un ensemble de règles (nommées "classeurs"), qui associent des conditions à une action (prédiction). Le tableau 1.2 présente un exemple de trois règles fictives issues de l'entrainement d'un LCS. Une règle (ou un classeur) se présente sous la forme "SI [condition] ALORS [prédiction]". Chaque règle est associée à un poids, ce qui permet de définir une importance plus ou moins forte. Leur fonctionnement est donc similaire aux arbres de décision, mais donc les règles sont évaluées sans ordre imposé.

Conditions	Prédiction	Poids
SI éternuement ET allergique au pollen	ALORS sain	7
SI temp. corporelle > 38°C	ALORS malade	12
SI éternuement ET temp. corporelle = 37°C	ALORS sain	3

Tableau 1.2 – **Exemple de règles fictives issues de l'entrainement d'un algorithme de LCS.** Les règles issues de l'entrainement d'un LCS sont sous la forme "SI [conditions] ALORS [prédiction]". À l'inverse de la forêt aléatoire où chaque arbre possède le même pouvoir de vote, chaque règle est associée à un poids, reflétant son importance lors du calcul de la prédiction finale.

L'apprentissage des systèmes de classeurs se réalise par des mécanismes d'évolution. Chaque règle va être initialement générée de façon aléatoire (ou guidée, R. J. Urbanowicz et al., 2018) puis modifiée (mutée) aléatoirement au fur et à mesure des cycles d'entrainement (générations). À chaque cycle, les règles les moins performantes lors de l'évaluation sur le jeu d'entrainement sont éliminées du jeu de règles. Ainsi au bout d'un certain nombre de cycles (générations), les règles générées qui ont survécu au processus de sélection sont performantes dans leurs tâches de classification.

Les LCS sont extrêmement explicables, car leur apprentissage génère une liste de règles parfaitement intelligible pour l'Homme, ainsi il est facile de reproduire le processus de prédiction des points de données manuellement (Arrieta et al., 2019). Aussi, pour chaque prédiction, il est possible de savoir exactement quelles règles ont été déclenchées et ont mené à cette prédiction. Cependant, ces systèmes restent sous-performants et leurs méthodes d'apprentissage par évolution posent des difficultés de mise à l'échelle et nécessitent un grand nombre de données pour être efficaces (R. J. Urbanowicz et Moore, 2015).

1.2.5 Limites du *machine-learning* appliqué aux données biomédicales

Bien que les techniques de ML se révèlent utiles pour traiter, analyser et prédire de grands ensembles de données, ces techniques font face à de grandes difficultés dans le cadre des données biomédicales (Martínez-García et Hernández-Lemus, 2022). Les données biomédicales étant massives, mais surtout non structurées et multimodales, il est difficile de réaliser le travail d'annotation nécessaire à leurs structurations pour l'utilisation d'algorithme de ML. Car l'annotation manuelle des données est un travail couteux en temps et en argent, d'autant plus dans le domaine médical ou l'annotateur doit être un expert du domaine. Ainsi il est nécessaire de développer des méthodes d'IA capables d'apprendre et d'exploiter les données brutes non structurées sous toutes les formes (textes, images, séquences).

Dans ce contexte, les réseaux de neurones profonds présentent une opportunité pour le traitement des données biomédicales non structurées. En effet, les réseaux de neurones sont parmi les modèles les plus performants et sont capables de traiter aisément des données non

structurées en extrayant eux-mêmes les descripteurs pertinents à partir de la structure des données. Cependant, ce sont aussi de complètes boites noires et non explicables. Une stratégie intéressante pour l'utilisation des réseaux de neurones profonds tout en gardant une explicabilité satisfaisante est de les coupler aux méthodes de ML classiques. Il est possible d'utiliser ces réseaux de neurones sous forme de boites noires pour extraire des descripteurs pertinents à partir de données non structurées (par exemple, quantifier un marqueur pathologique sur une image). Puis dans un second temps, d'utiliser des méthodes de ML classiques et explicables pour, à partir de ces descripteurs extraits, réaliser la tâche de classification et de diagnostic.

Dans le prochain chapitre, nous verrons comment les réseaux de neurones profonds fonctionnent et peuvent être utilisés pour traiter et extraire de l'information des données biomédicales non structurées là où les techniques de ML classique échouent.

Chapitre 2

Réseaux de neurones et traitement de données biomédicales non structurées

> "People should stop training radiologists now. It's just completely obvious that in five years deep learning is going to do better than radiologists."
>
> – Geoffrey Hinton, 2016

La vision de Geoffrey Hinton, lauréat du prix Turing 2018 pour ses travaux en IA et en réseaux de neurones profonds (*Deep Neural Networks*, DNN), était peut-être un petit peu optimiste. Presque huit ans après cette prédiction, les radiologues n'ont pas été remplacés par l'IA et continuent d'être formés. Cependant, il est important de remarquer que les méthodes IA ont fait des progrès considérables et peuvent présenter des performances similaires aux radiologues, par exemple dans l'évaluation de radiographies de poumon (Frauke Rudolf, 2023). En juin 2023, il y a 238 produits médicaux basés sur IA pour la radiologie et autres méthodes d'imageries avec autorisation de mise sur le marché par la *Food and Drug Administration (FDA)* (Keith J. Dreyer et al., 2023). Si l'IA n'est pas encore prête à remplacer les radiologues et praticiens dans l'évaluation des données d'imagerie, elle est au moins maintenant capable de les assister afin de permettre un gain de temps et de précision dans l'évaluation des données.

Au cours de la dernière décennie, le domaine de l'IA a été révolutionné par l'apparition des réseaux de neurones profonds (*Deep Neural Networks*, DNN) grâce notamment aux travaux des Yann Lecun, Geoffrey Hinton et Yoshua Bengio. Cette nouvelle technologie d'IA fait une promesse intéressante dans le cadre des données biomédicales : être capable de traiter automatiquement des données non structurées, c'est-à-dire sans devoir définir des descripteurs pertinents manuellement. Il est donc tout à fait pertinent d'explorer comment ces modèles

peuvent être exploités pour le traitement des données biomédicales multimodales et hétérogènes. Dans ce chapitre, nous allons d'abord présenter le fonctionnement des réseaux de neurones. Puis nous allons étudier deux architectures spécifiques de réseaux de neurones qui ont permis la création de modèles d'analyse d'images (réseaux convolutifs) et de texte libres (réseaux de types *transformers*).

2.1 Réseaux de neurones profonds

2.1.1 Le concept de neurones et réseaux de neurones profonds

Les réseaux de neurones sont un concept ancien qui a été décrit pour la première fois en 1958 par Frank Rosenblatt (Rosenblatt, 1958) sous sa plus simple forme nommée le perceptron, un réseau composé d'un seul neurone formel. Les réseaux de neurones reposent sur le concept bio-inspiré des neurones. La figure 2.1 présente les similarités entre un neurone biologique et un neurone formel en IA. Le neurone biologique, par des processus biochimiques, capte les signaux d'entrée de l'environnement ou des neurones précédents par les dendrites. Ces signaux sont intégrés dans le corps cellulaire pour transmettre ou non un signal de sortie à travers l'axone vers d'autres neurones. Le neurone formel d'IA est un modèle simplifié du neurone biologique qui mime leur fonctionnement. Ainsi le neurone formel intègre des entrées (x1, x2, x3 sur le schéma) comme les dendrites, il calcule un signal à transmettre (par la somme pondérée des entrées et la fonction d'activation) comme le corps cellulaire et transmet ce signal aux neurones suivants (sortie) comme les axones.

Le perceptron, réseau de neurones composé d'une seule couche d'un ou plusieurs neurones dans sa version plus avancée entre l'entrée et la sortie, n'est efficace que pour traiter des problèmes à séparation linéaire. Pour traiter des problèmes de classification plus complexes, il est nécessaire de multiplier les couches de neurones entre l'entrée et la sortie du réseau. Ces couches sont appelées couches "cachées", et forment ce qu'on appelle un réseau de neurones profond. La limite entre réseaux de neurones classiques (perceptron multi-couche) et réseaux de neurones profonds est floue. Certains auteurs peuvent considérer un réseau de neurones comme profond à partir de 3 couches cachées, pour d'autres, entre 10 couches et 100 sont nécessaires.

À l'instar du cerveau, les neurones formels (artificiels) sont présents en grand nombre dans les réseaux de neurones profonds et sont interconnectés selon une organisation précise, cette organisation se nomme l'architecture du réseau.

2.1.2 Les réseaux de neurones : une diversité d'architectures

La multiplication du nombre des neurones dans un réseau de neurones et de leur interconnexion permet de constituer des architectures spécifiques qui confèrent des compétences particulières au réseau de neurones comme l'intégration d'informations locales (pour l'imagerie par exemple grâce à la convolution, Fukushima, 1980), l'intégration d'informations globales (pour l'analyse de séquences comme des phrases par exemple grâce aux *transformers*, Vaswani et al., 2017), des mécanismes de "mémoires" (pour l'analyse de textes par exemple grâce à la *Long short-term memory*, LSTM, Hochreiter et Schmidhuber, 1997) ou encore des capacités génératives (pour la génération d'images par exemple grâce aux réseaux antagonistes génératifs (GAN) et aux modèles de diffusion). La figure 2.2 (Leijnen, 2016) présente graphiquement quelques architectures de réseaux de neurones communes. On y retrouve le perceptron (P), le perceptron multi-couche (DFF), le réseau LSTM, le DNN convolutif (DCN) et les réseaux GAN.

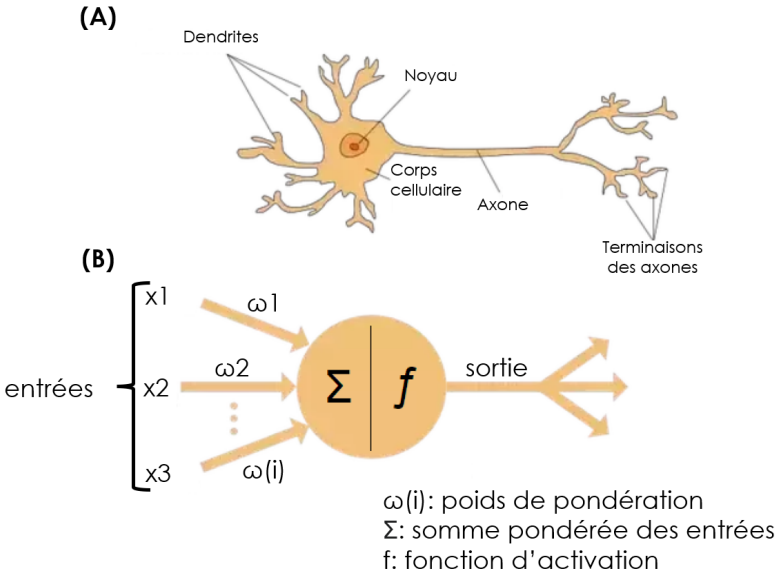


FIGURE 2.1 – **Comparaison du neurone biologique et neurone formel. (A)** Représentation schématique du neurone biologique. **(B)** Représentation schématique du neurone formel utilisé en IA.

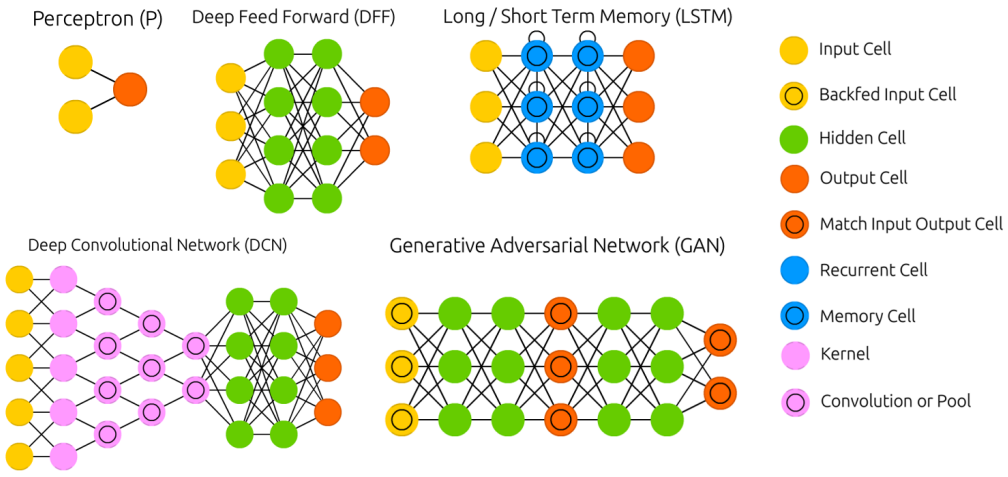


FIGURE 2.2 – **Représentation de différentes architectures de DNN.** Le type de neurones, leurs organisations en couches et les interconnections qui les lient définissent des architectures qui leur confère certaines capacités comme le traitement d'images (DCN) où encore la mémoire à court terme (LSTM). (modifié de LEIJNEN, 2016)

2.1.3 L'entrainement d'un réseau de neurones

L'entrainement d'un réseau de neurones consiste à trouver, pour chaque neurone qui le compose, la valeur de poids (ω) pour chaque entrée optimale pour remplir la tâche qui lui est assignée. Pour cela un réseau de neurones a besoin de deux éléments : (i) une fonction de cout permettant d'évaluer son niveau d'erreur de classification et (ii) une méthode qui permet de modifier les poids des connexions neuronales pour réduire cette erreur grâce à la descente de gradient et à la rétropropagation.

2.1.3.1 Fonction de cout

La fonction de cout permet de fournir une mesure quantitative des performances d'un réseau de neurones profonds. Elle mesure la divergence entre les prédictions réalisées par le modèle et les labels véritables des données. Cette fonction varie en fonction de la tâche à réaliser, pour une tâche de régression, un choix courant est l'utilisation de l'erreur quadratique moyenne (*mean square error*, MSE). Pour une classification, il est commun d'utiliser l'entropie croisée binaire (*binary cross entropy*).

Dans tous les cas, plus cette valeur est élevée, plus les prédictions du modèle divergent de la vérité de terrain, plus cette valeur est proche de 0, plus les prédictions du modèle sont exactes. Ainsi, l'objectif de l'entrainement d'un réseau de neurones est de faire converger la valeur de la fonction de cout du jeu d'entrainement et du jeu de validation vers 0.

La figure 2.3 présente un exemple théorique de valeurs de fonction de cout au cours de l'entrainement d'un modèle DNN. Dans cet exemple, l'écart tout au long de l'apprentissage entre le jeu de validation et d'entrainement est important à des fins de visualisation, en pratique, les deux courbes doivent presque se superposer. On observe ici qu'au cours de l'entrainement cette valeur converge vers 0. Cependant, au bout d'un moment, la valeur pour le jeu d'évaluation remonte tandis que celle du jeu d'entrainement continue de baisser, il s'agit du phénomène de sur-apprentissage, le modèle cesse d'apprendre à généraliser et apprend simplement par cœur le jeu d'entrainement. Dès lors, il est nécessaire d'arrêter l'apprentissage au moment où la valeur du jeu de validation augmente, il s'agit de ce qu'on appelle l'arrêt prématuré pour éviter le sur-apprentissage.

2.1.3.2 Rétropropagation et descente de gradient

La figure 2.4 (Scalzitti, 2021) présente le mécanisme de rétropropagation. Lors d'une prédiction, les signaux sont propagés vers l'avant (nommé *forward pass*) à partir de la couche d'entrée jusqu'à la couche de sortie, puis l'erreur est ensuite calculée grâce à la fonction de cout. Lors de l'entrainement, le chemin inverse est réalisé en propageant le gradient de l'erreur pour identifier les neurones responsables des erreurs (*backward pass*) (LeCun et al., 2015). Ce processus de rétropropagation permet d'identifier les neurones responsables des erreurs dont les paramètres doivent être modifiés pour réduire la fonction de cout.

Après avoir identifié les neurones à modifier (et donc avoir calculé le gradient de chaque neurone), leurs poids vont être ajustés grâce à la méthode de la descente de gradient. La figure 2.5 présente le fonctionnement de la descente de gradient. À chaque étape d'apprentissage, les poids des neurones (ici un seul poids pour un neurone sur la figure) sont mis à jour dans la direction négative du gradient, ce qui a pour effet de réduire la valeur de la fonction de cout. Cette modification des poids est proportionnelle à un paramètre nommé le pas d'apprentissage (*learning rate*). Ces cycles d'apprentissage vont se répéter jusqu'à atteindre un minimum (global ou local) dans la fonction de cout, indiquant un poids optimal pour la fonction de cout (et donc la tâche à réaliser).

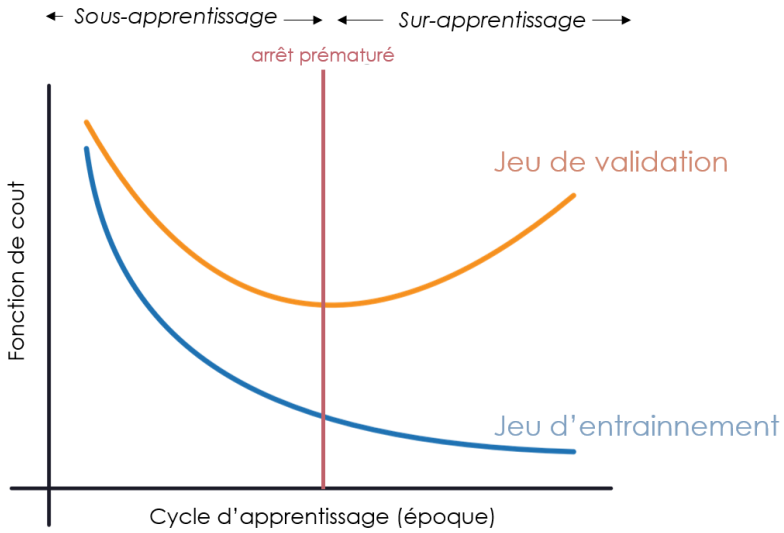


FIGURE 2.3 – **Schéma d'un exemple de fonction de coût lors de l'entrainement d'un DNN.** Au cours de l'entrainement, la fonction de cout baisse après chaque cycle d'apprentissage à la fois pour le jeu d'entrainement et le jeu de validation. Puis celle du jeu d'entrainement continu de baisser tandis que celle du jeu de validation augmente : le modèle commence un sur-apprentissage, il est nécessaire de stopper l'entrainement.

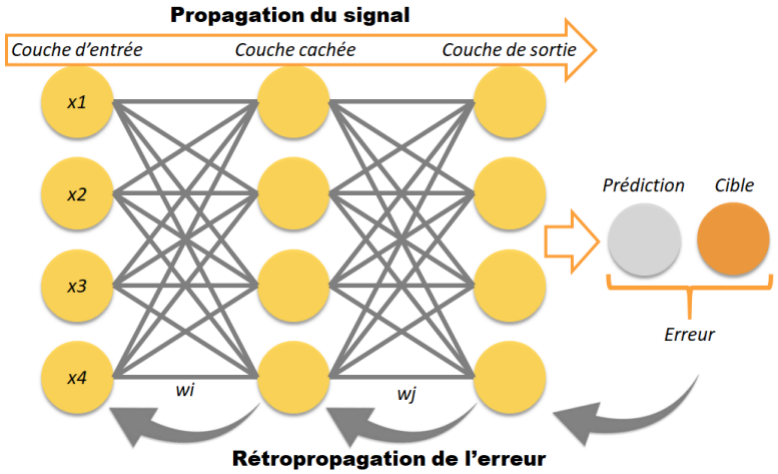


FIGURE 2.4 – **Schéma de la propagation du signal lors de la prédiction et de la rétropropagation lors de l'entrainement.** Lors d'une prédiction, le signal est propagé vers l'avant à travers les neurones, puis une erreur est calculée. Cette erreur est ensuite propagée dans le sens inverse à travers les neurones, c'est la rétropropagation. (SCALZITTI, 2021)

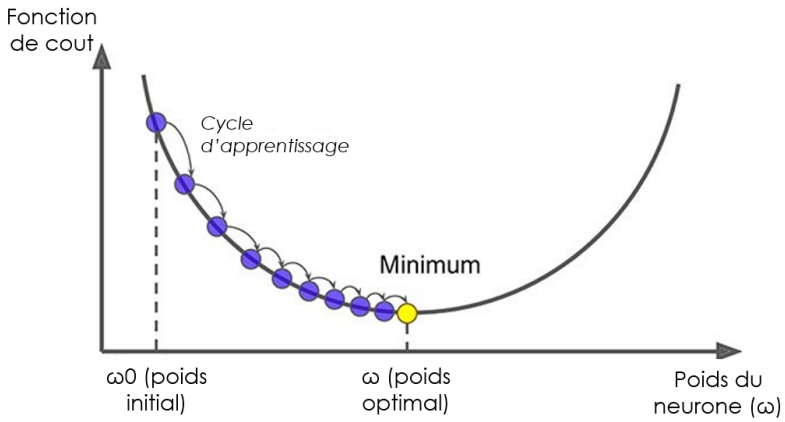


FIGURE 2.5 – **Schéma de la descente de gradient pour un poids d'un neurone**. À chaque cycle d'apprentissage, le poids du neurone est mis à jour jusqu'à un minimum local qui minimise la fonction de cout.

La rétropropagation et la descente de gradient permettent l'apprentissage de réseaux de neurones composés de centaines de milliers, voire de millions et même de milliards de neurones. La taille des DNN est très variable en fonction des architectures et de la tâche à effectuer. Ainsi, les couts en matière de calcul pour leur entrainement (le calcul des gradients et la mise à jour des poids) peuvent devenir très importants. Dans la prochaine sous-section, nous allons présenter des exemples de réseaux et d'architectures de DNN et les ressources informatiques associées nécessaires à leur utilisation.

2.1.4 Nombre de paramètres et ressources informatiques

Il existe un lien de proportionnalité entre la taille (en nombre de neurones) d'un DNN, sa précision et les ressources de calcul nécessaires à son entrainement et aux prédictions. En théorie, plus un réseau est grand, plus importante est sa capacité à capturer des relations complexes entre les données et meilleure est sa précision. Le tableau 2.1 (CHOLLET, 2023) présente plusieurs architectures utilisées pour la classification d'images. On observe que pour une même architecture, augmenter le nombre de paramètres permet d'obtenir de meilleures performances, mais au détriment d'un temps d'inférence plus long. Cependant, lors d'une comparaison de deux architectures différentes, la relation performance - nombre de paramètres n'est pas forcément vérifiée. Les réseaux EfficientNet (architecture de classification d'image développée en 2020 pour améliorer les ResNet TAN et LE, 2020) performent mieux que les réseaux ResNet (architecture de 2015 HE et al., 2015) avec un nombre de paramètres moindre, cependant une règle qui reste vérifiée dans toutes les conditions : les réseaux qui performent le mieux ont un temps d'inférence plus long, quelle que soit leur architecture.

L'entrainement et l'inférence des DNN se réalise sur un matériel informatique spécialisé nommé carte graphique (*Graphical Processing Unit*, GPU). Au contraire des processeurs centraux (CPU) qui sont conçus pour des opérations séquentielles, les GPU permettent de réaliser un grand nombre d'opérations mathématiques en parallèle. Or, les opérations principales nécessaires pour l'entrainement et l'inférence d'un DNN sont des calculs matriciels et donc intrinsèquement parallélisables. La caractéristique limitante des GPU concerne leur mémoire disponible (VRAM). En effet, plus le modèle de DNN est grand, plus il possède de paramètres, plus il va être demandeur en mémoire, car il est nécessaire de pouvoir garder l'ensemble des poids à optimiser

Architecture	Paramètres (millions)	Précision (ImageNet, %)	Temps d'inférence (ms)
MobileNetV2	3,5	71,3	3,8
ResNet50V2	25,6	74,9	4,6
ResNet101V2	44,7	77,2	5,4
ResNet152V2	60,4	78,0	6,6
EfficientNetB1	7,9	79,1	5,6
EfficientNetB2	9,2	80,1	6,5
EfficientNetB3	12,3	81,6	8,8

Tableau 2.1 – **Tableau de comparaison d'architectures de DNN et performances**. Pour un même type d'architecture, l'augmentation du nombre de paramètres améliore les performances, mais réduit aussi la vitesse d'inférence. Cependant un nombre de paramètres plus élevé ne signifie pas forcément de meilleures performances quand on compare deux architectures différentes. (Chollet, 2023)

en mémoire.

Ainsi, s'il est possible d'entrainer et de faire de l'inférence de modèles de taille raisonnable sur des GPU accessibles au grand public, cette tâche devient complexe, voire impossible, pour des modèles de très grande taille, sans engager des couts de plusieurs centaines de milliers d'euros de matériel. À titre d'exemple, il est possible d'héberger et d'entrainer des modèles de quelques dizaines de millions de paramètres à quelques milliards de paramètres (tel que Resnet50 (25 millions de paramètres) et LLaMA-7B (7 milliards de paramètres) sur un seul GPU grand public (16 à 24 Go de mémoire).

Parmi les modèles les plus grands à ce jour, on retrouve les modèles développés pour l'analyse de langages (modèles linguistiques de grande taille (*Large Language Models*, LLMs)) présentés dans une prochaine section (2.3.5). Ces modèles atteignent une taille de plusieurs dizaines voire des centaines de milliards de paramètres : par exemple GPT-3 d'OpenAI (175 milliards de paramètres) (Brown et al., 2020) ou LLaMA de META (65 milliards de paramètres) (Touvron et al., 2023). Ce type de modèles demande l'utilisation de plusieurs GPU haut de gamme (4 à 8) pour leur hébergement et inférence. À titre indicatif, un cluster composé de 4 GPU H100 (40,000€ pièce), coute 17.9$/heure d'utilisation, soit environ 13 000$ d'hébergement mensuel pour un modèle accessible en continu. Un travail d'optimisation pour réduire la taille des modèles tout en conservant leurs performances est donc nécessaire pour permettre d'utiliser ce type d'architecture et de dépasser les challenges liés à leur hébergement.

2.2 L'analyse d'imagerie et de séquences par réseau neuronal convolutif

L'architecture des réseaux de neurones convolutifs (CNN) permet l'analyse et l'exploitation de données de type image brute. Cette architecture est majoritairement utilisée pour la classification et segmentation d'images, mais elle est aussi capable d'exploiter des données de types séquences (phrases, séquences génomiques...). Elle est basée sur les travaux réalisés sur le cortex visuel de chats et de primates dans lesquels il a été démontré que chaque neurone du cortex visuel captait une information locale, dans un champ visuel réduit. De plus, chaque neurone n'est en mesure de capter qu'une orientation de ligne (horizontale, verticale ou oblique) (Hubel, 1959; Hubel et Wiesel, 1959). À partir de ces travaux, le concept de couches convolutives pour les réseaux de neurones a été formulé (Fukushima, 1980) et a permis la construction du premier

DNN convolutif pour reconnaitre des numéros sur des chèques de banque grâce au réseau LeNet-5 (LeCun et al., 1998). Vingt-cinq ans plus tard, en 2023, les réseau neuronal convolutif (*Convolutional Neural Networks*, CNN) sont encore utilisés pour l'analyse de données d'imagerie biomédicale (Hölscher et al., 2023; Ker et al., 2019).

2.2.1 Fonctionnement des couches convolutives pour l'analyse d'images

Pour simuler le comportement des neurones du cortex visuel, les neurones d'une couche convolutive ne sont connectés qu'à une zone restreinte d'une image, généralement sous forme d'un carré de pixels. La figure 2.6 (Aurélien Géron, 2019) présente schématiquement la liaison de trois neurones répartis sur deux couches convolutives par rapport à une image de base. On observe que ces neurones n'ont accès qu'à une portion de l'image, donc à une information locale.

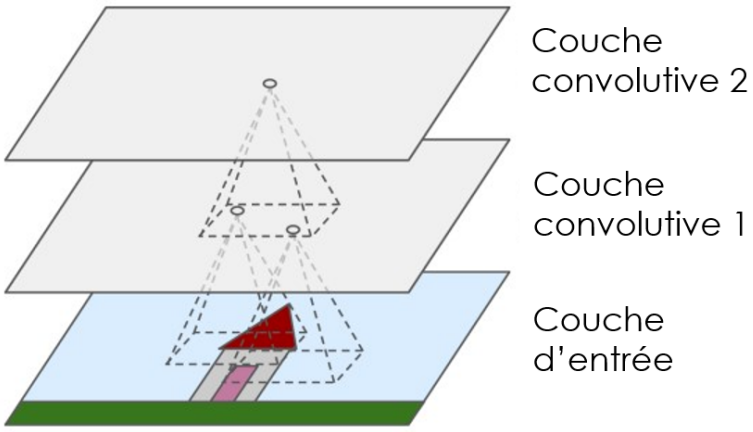


Figure 2.6 – **Schéma de la connexion des neurones convolutifs à une image.** Les neurones ne sont connectés qu'à une portion locale de l'image sous la forme d'un carré de pixels. (modifiée de Aurélien Géron, 2019)

La convolution consiste à appliquer un filtre à une entrée pour produire une carte de caractéristiques (*feature map*). Ainsi les neurones de la couche convolutive ont en entrée une portion de l'image et vont appliquer un filtre pour extraire une information de cette portion (une ligne horizontale, une texture, un contraste...). Le but de l'entrainement du CNN est, pour chaque neurone, de trouver les filtres optimaux à utiliser pour extraire l'information pertinente à la classification de l'image. Autrement dit, chaque neurone va apprendre à extraire une information pertinente de la zone à laquelle il a accès.

La figure 2.6 (Aurélien Géron, 2019) est simpliste, car elle représente les couches convolutives comme composées d'une seule couche de neurones et donc un seul filtre. En réalité, comme présentée en figure 2.7 (Aurélien Géron, 2019), chaque couche convolutive est composée de plusieurs filtres (couches de neurones formant des cartes de caractéristique). Chacune de ces cartes de caractéristique (*feature map*) est reliée aux précédentes afin d'extraire des caractéristiques très diverses (des formes horizontales, verticales, de la texture, du contraste...).

Enfin pour simplifier les couts de calcul et réduire la mémoire nécessaire, après un bloc de convolution, une étape de *max-pooling* (en français : sous-échantillonnage maximal) est réalisée. Cette technique permet de réduire la dimensionnalité des cartes de caractéristiques en préservant les informations essentielles. La figure 2.8 (Aurélien Géron, 2019) présente le fonctionnement de cette méthode. La carte de caractéristique est divisée en carré non chevauchant et la valeur

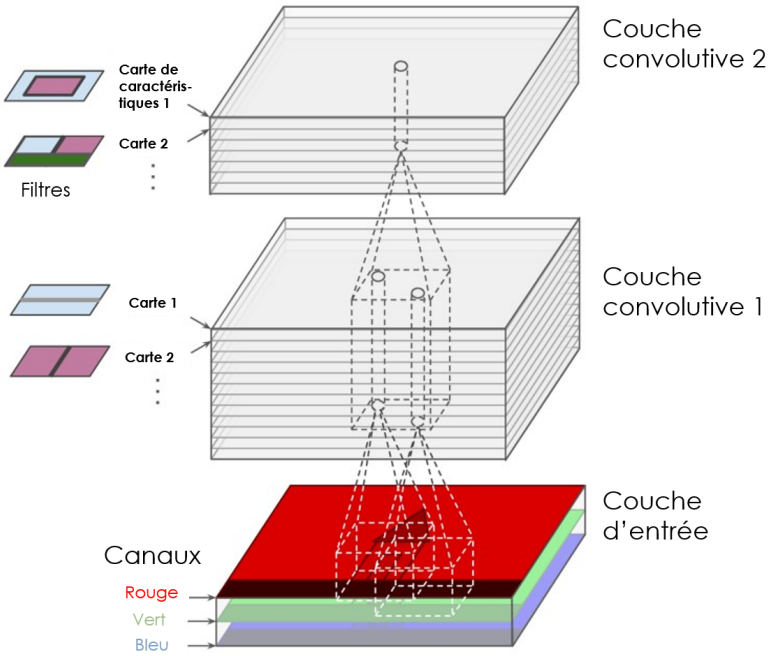


FIGURE 2.7 – **Schéma des couches convolutives complexes**. Chaque couche convolutive est en réalité composée de plusieurs filtres, chacun générant une carte de caractéristique propre. (modifiée de AURÉLIEN GÉRON, 2019)

maximale de chaque carré est sélectionnée. Par exemple, si on a une carte de caractéristique de taille 4x4 et qu'on réalise un *max pooling* de taille 2, on obtient une carte de caractéristique de taille 2x2, c'est-à-dire quatre fois plus petite.

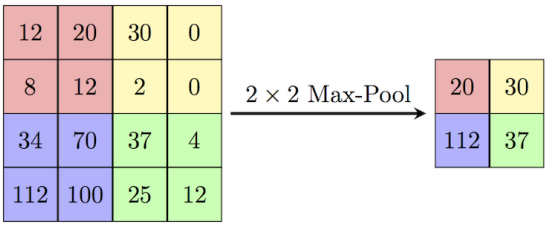


FIGURE 2.8 – **Schéma de la technique de *max-pooling* pour réduire la dimensionnalité d'une matrice**. Le max-pooling permet de réduire la dimension d'une carte de caractéristique en conservant les informations importantes, réduisant ainsi les couts de calcul. (AURÉLIEN GÉRON, 2019)

Pour finir, la figure 2.9 (AURÉLIEN GÉRON, 2019) présente la structure typique d'un CNN. Un CNN consiste en l'enchainement de couches de convolution et de *max-pooling* avant les couches de classification (*fully connected*, une couche de neurones où tous les neurones sont connectés à la couche précédente.). Le but de cette architecture est de réduire la dimensionnalité de l'image tout en augmentant la profondeur (c'est-à-dire le nombre d'informations extraites par position). Par exemple, pour une image de 512x512 (c'est-à-dire 262 144 pixels), si en sortie de couches convolutives on obtient une matrice de taille (4x4x128, soit 2048 points), cela signifie que nous avons extrait 128 caractéristiques pour chacune des 16 zones de l'image. Ces 128 caractéristiques par zone vont être utilisées pour classer l'image.

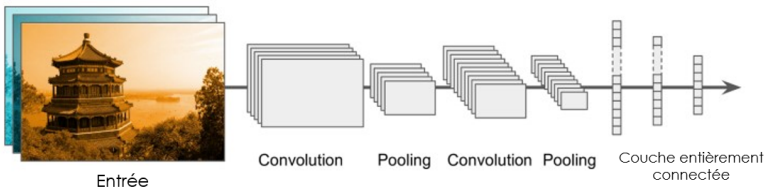


FIGURE 2.9 – **Schéma de la structure d'un CNN typique.** Une architecture de CNN typique consiste en un enchaine de couche convolutives et de max-pooling (extraction d'information), puis d'une couche finale entièrement connectée pour la classification (modifiée de AURÉLIEN GÉRON, 2019)

2.2.2 Modèle généraliste pour l'histologie

La couche de neurones convolutifs représente la brique de base des CNN. À partir de cette architecture de réseau, il est possible de réaliser diverses tâches en variant légèrement l'organisation des couches. Dans cette section, nous présentons deux exemples d'utilisation des CNN pour le traitement de données biomédicales.

2.2.2.1 Segmentation d'images histologiques : détection de cellules avec Cellpose

La segmentation d'une image consiste à diviser l'image en groupes de pixels (segment) dans l'objectif d'obtenir les coordonnées d'éléments d'intérêts. Par exemple, dans le cadre d'une image de coupe histologique (image de tissus au microscope), il peut être intéressant de mesurer le nombre de cellules présentes et leur taille. Pour automatiser ce processus, il est nécessaire d'avoir recours à un réseau de neurones capable de réaliser de la segmentation d'image.

Cellpose (Stringer et al., 2021), développé par Carsen Stringer en 2021, est un modèle de segmentation généraliste, conçu pour être capable de segmenter les cellules de n'importe quelle coupe histologique. La figure 2.10 présente l'architecture du modèle ainsi qu'un exemple d'image histologique et de résultat de segmentation. L'architecture de CNN utilisée se nomme U-Net (Ronneberger et al., 2015) structurée comme un U avec un chemin de contraction (encodeur, grâce à la convolution) et un chemin d'expansion (décodeur grâce à la "*upconvolution*"). Cette architecture permet à partir de l'image d'entrée, composée de cellules en microscopie à fluorescence, de générer un masque de segmentation, de la même taille que l'image d'entrée. Ce masque de segmentation est une abstraction de l'image d'entrée où les pixels de chaque objet (cellules) sont marqués avec un identifiant unique. Ainsi il est possible à partir de ce masque de compter ou de mesurer la taille des cellules.

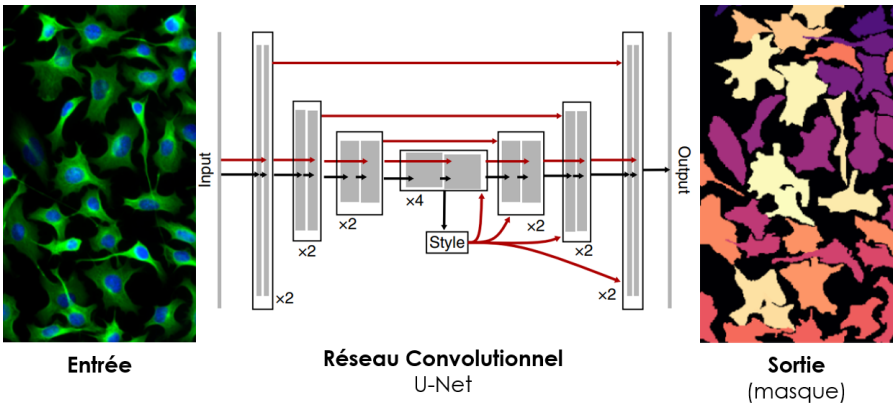


Figure 2.10 – **Architecture du réseau convolutif de CellPose**. Cellpose est composé d'un réseau CNN de type U-Net. Cette architecture en forme de U est composée d'un chemin de contraction (un encodeur grâce à la convolution qui extrait des informations de l'image) puis d'un chemin d'expansion (décodeur grâce à l'*upconvolution*) qui reconstitue une image simplifiée, sous la forme d'un masque de segmentation. (modifié de Stringer et al., 2021)

2.2.2.2 Analyse de séquences : prédiction de sites d'épissage avec Spliceator

En plus du traitement d'images, les CNN peuvent être utilisés pour traiter des données de type séquences nucléotidiques. Au sein de notre équipe, en 2021, l'outil Spliceator a été développé (Scalzitti et al., 2021) pour analyser des séquences génomiques et prédire les sites d'épissages. La figure 2.11 présente la structure du CNN entrainé pour cette tâche de classification. Le CNN entrainé est capable de prédire les sites d'épissage de plus de 100 espèces vivantes avec une précision de plus de 90%. L'architecture utilisée par Spliceator est une architecture classique de CNN avec trois blocs convolutifs puis une couche *fully-connected* pour la classification, identique à l'exemple en figure 2.9. Cette architecture a permis d'entrainer un modèle capable de prendre en compte un contexte jusqu'à 600 nucléotides en amont et en aval du site d'épissage à évaluer.

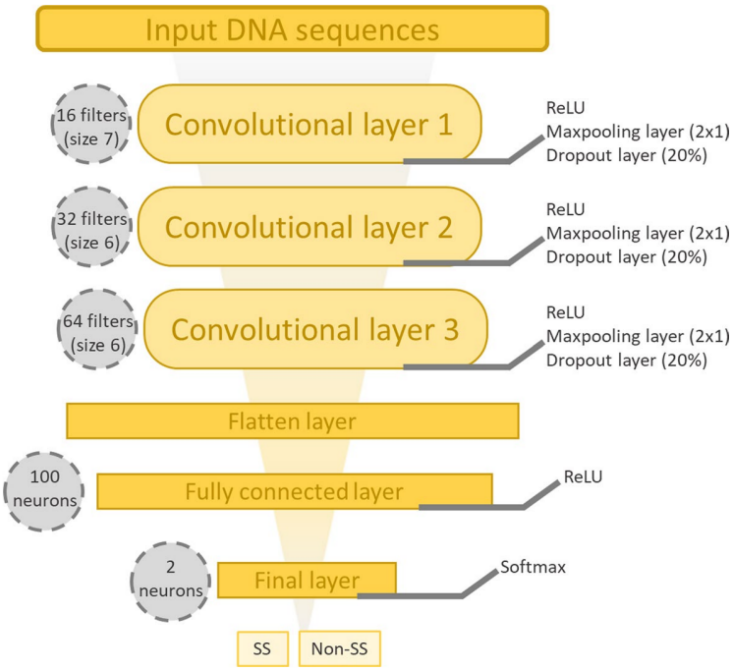


FIGURE 2.11 – **Architecture du réseau convolutif de Spliceator**. Une suite de trois couches convolutives sont utilisées pour extraire les caractéristiques qui constituent les sites d'épissages puis une couche entièrement connectée utilise ces caractéristiques pour apprendre à reconnaitre les sites d'épissages. (SCALZITTI et al., 2021)

Bien que les CNN soient capables d'intégrer de l'information locale pour mimer le fonctionnement du contexte visuel pour une analyse d'image, la détection et l'intégration d'interactions longue distance demeurent un problème majeur. Cette limite pose problème dans le cadre de l'analyse de séquences. Par exemple, si l'on cherche à prédire l'expression d'un gène, il existe des signaux à longue distance qui modulent cette expression, tels que les séquences *enhancers*, dont plus de la moitié sont situées à des distances de 50 000 paires de bases ou plus du gène modulé (Chepelev et al., 2012). De même pour l'analyse de texte, une information importante pour la compréhension d'une phrase peut être située bien en amont à plusieurs dizaines, voir des centaines de mots.

Ainsi pour construire des architectures de DNN capables de s'attaquer aux défis de l'analyse de textes et de séquences, il est nécessaire d'avoir des modules capables de prendre en compte le contexte global. Ceci est rendu possible grâce à l'architecture nommée *transformer* (Vaswani et al., 2017), qui est présentée dans la prochaine section.

2.3 Architecture transformer et la révolution des modèles linguistiques de grande taille

Les *transformers* sont des architectures qui ont révolutionné le domaine de l'analyse de séquence et de texte depuis leur première description (Vaswani et al., 2017). Cette architecture a permis de dépasser toutes les limitations des architectures précédentes, en particulier les limites concernant leur faible mémoire à longue distance. Les modèles *transformers* sont des modèles dits de séquence à séquence (*Seq2Seq*), qui prennent en entrée une séquence (une phrase) et produisent en sortie une autre séquence (par exemple la phrase traduite). Pour comprendre comment les *transformers* fonctionnent et quelles sont leurs implications, deux notions sont nécessaires : la structure encodeur-décodeur et les mécanismes d'attention.

2.3.1 La structure encodeur-décodeur et l'attention multitête

La structure encodeur-décodeur est au cœur de nombreuses architectures et est vitale pour une multitude de tâches telles que la génération de texte ou la traduction. L'idée derrière cette structure est de compresser l'information d'entrée en un vecteur numérique de taille définie, nommé "état caché" (figure 2.12). Puis le décodeur, à partir de cet état caché, va générer une séquence (par exemple une phrase) qui correspond à cet état caché. Lors de l'entrainement d'un réseau encodeur-décodeur, les paramètres de l'encodeur et du décodeur sont ajustés pour minimiser la fonction de cout.

Par exemple, imaginons que l'on veuille construire un réseau de traduction du français vers l'anglais (figure 2.12). Pour cela, on va utiliser comme données d'entrainement des couples de phrases françaises (entrée) et anglaises (cible). L'entrainement consiste à apprendre à l'encodeur à représenter numériquement les phrases françaises en extrayant les informations les plus pertinentes. L'entrainement du décodeur consiste à apprendre comment à partir de cet état caché, reconstruire la phrase anglaise attendue.

Cependant, cette structure possède des limites. En effet, pour des phrases très longues ou des paragraphes, l'encodeur n'est pas nécessairement en mesure de représenter toutes les informations pertinentes dans l'état caché qui possède une taille fixe, ainsi de l'information pourrait être perdue. Les mécanismes d'attention à l'œuvre dans les *transformers* permettent de pallier cette limite.

Dans un *transformer*, les encodeurs et les décodeurs sont multiples (au total de 6 dans le papier les décrivant pour la première fois) et ils sont composés de mécanismes d'auto-attention

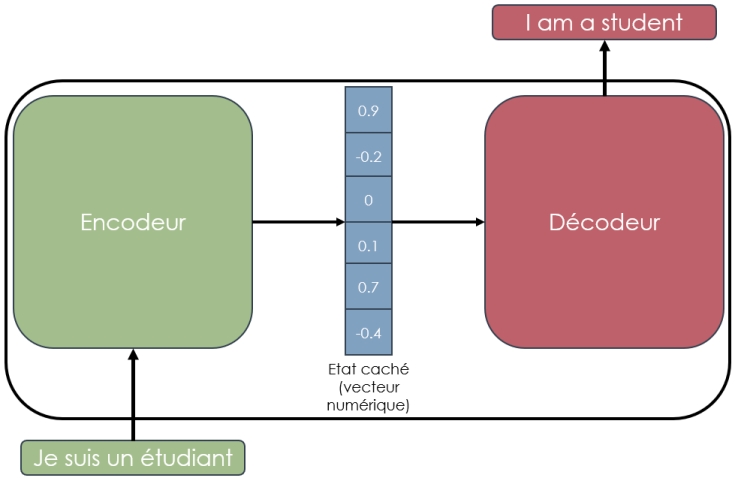


FIGURE 2.12 – **Schéma de la structure encodeur-décodeur**. Cette structure permet l'encodage d'une donnée dans un état caché puis son décodage pour la traduction par exemple.

(figure 2.13). Le mécanisme d'auto-attention permet à n'importe quel élément de la séquence (mot de la phrase) d'être influencé par les autres éléments. C'est-à-dire que le premier mot d'une phrase peut avoir une influence sur le dernier, ce qui permet une influence longue distance. Mathématiquement, cela est possible par la création pour chaque mot de la phrase de trois vecteurs : requête (Q), clé (K) et valeur (V) (non représenté sur la figure 2.13). Puis par un produit scalaire, un score d'attention est calculé pour tous les couples de mots présents dans la phrase (c'est donc un processus quadratique).

Ces mécanismes d'attention sont divisés en plusieurs têtes, qui permettent lors de l'entrainement de spécialiser chacune des têtes d'attention sur un aspect spécifique de l'entrée. Par exemple, dans le cadre d'un texte, une tête peut être spécialisée dans la syntaxe et une autre dans des aspects sémantiques. La multiplicité des *transformers* et les mécanismes d'attention multitête permettent aux réseaux de traiter des données de grandes tailles sans perte d'informations, c'est-à-dire avec une grande "mémoire". Ainsi, cela permet de détecter et d'intégrer lors des prédictions réalisées des informations à longue distance, là où la convolution ne permettait d'intégrer que des informations locales.

Grâce à cette structure encodeur-décodeur et aux mécanismes d'attention multitête, il est possible de créer une architecture de *transformer* présentée en figure 2.14. Cette architecture est composée de deux éléments : à gauche en vert, les encodeurs empilés au nombre de 6, qui prennent en entrée la phrase à traduire. À droite en rouge, les décodeurs empilés au nombre de six prennent à la fois la sortie des encodeurs en compte ainsi que la traduction dont la génération a déjà commencé. Dans chaque décodeur et encodeur, on retrouve les modules d'attentions (en bleu). L'inférence de cette architecture est une boucle (flèche violette), les mots sont générés un par un à chaque cycle. Ainsi pour la traduction en anglais de la phrase "Je suis un étudiant", le mot "I" sera généré au premier cycle et sera ensuite repris en entrée des décodeurs pour générer le suivant.

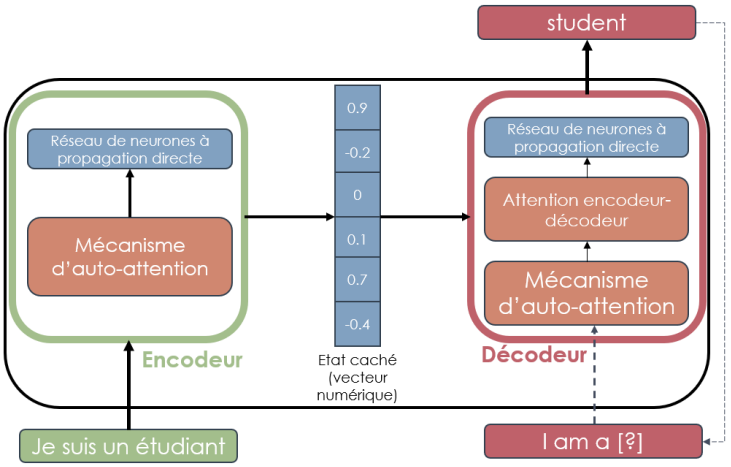


FIGURE 2.13 – **Schéma de la structure encodeur-décodeur d'un réseau transformer**. Chaque encodeur et décodeur est composé d'un mécanisme d'attention et d'un réseau à propagation directe.

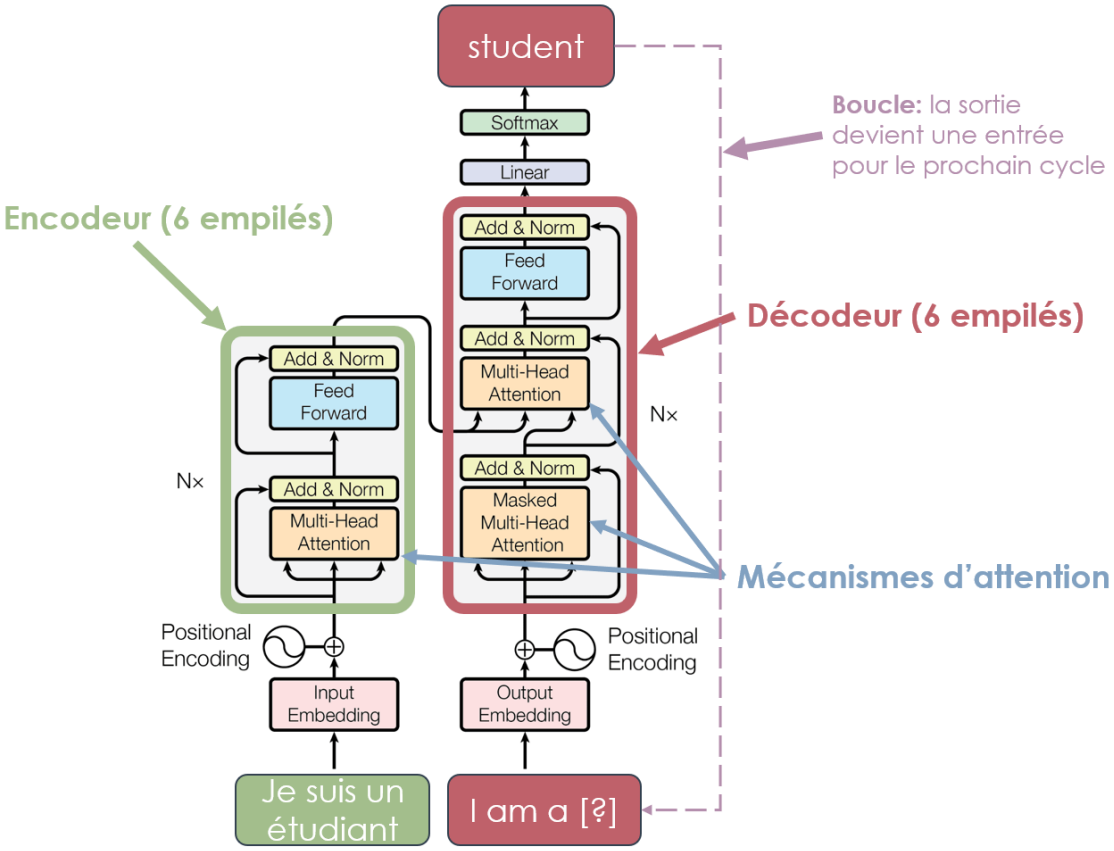


FIGURE 2.14 – **Schéma de l'architecture d'un réseau transformer.** Les transformer sont composés de 6 encodeurs empilés (en vert), reliés à 6 décodeurs empilés (en rouge), chacun composé de mécanismes d'attention (en bleu). Les prédictions du modèle se font suivant une boucle (en violet) où à chaque étape, la prédiction actuelle est utilisée comme entrée pour la prédiction suivante.

Ici pour expliquer le fonctionnement des *transformers*, nous avons utilisé l'exemple de la traduction de texte. Mais les *transformers* peuvent être utilisable pour n'importe quelles données séquentielles, telles que les séquences génomiques et protéiques. Il est possible de considérer les séquences d'ADN et de protéine comme un simple enchainement de lettres, un nouveau langage à traduire. Nous allons voir deux exemples d'utilisation de *transformers* pour l'analyse de séquences biologiques à travers Enformer et AlphaFold.

2.3.2 Enformer : un *transformer* pour l'expression des gènes

Comme écrit précédemment, l'expression des gènes dans le génome est modulée par des interactions longue distance, grâce notamment aux *enhancers*. Ainsi, si l'on veut construire des modèles prédictifs de l'expression des gènes, il faut avoir une architecture capable de considérer ces interactions longue distance, c'est précisément l'intérêt des *transformers*. Dans les travaux intitulés "*Effective gene expression prediction from sequence by integrating long-range interactions*" (Avsec et al., 2021) publié en 2021, une équipe de DeepMind a utilisé un modèle mixant convolutions et *transformers* pour prédire l'expression des gènes (figure 2.15. Là où le précédent modèle n'était capable de prendre que 20 000 paires de bases de contexte autour d'un gène pour effectuer sa prédiction, ce nouveau modèle nommé *Enformer* est capable d'étendre le contexte jusqu'à 100 000 paires de bases en amont et en aval du gène. Grâce à l'examen des prédictions réalisées, ils ont réussi à mettre en évidence que le modèle a porté son attention sur la présence d'*enhancers* localisés à 50 000 paires de bases du gène. Ce qui confirme l'intérêt de l'utilisation de *transformers* dans le cadre de l'analyse de l'expression des gènes et de la prise en compte des *enhancers*. Cependant, de récents travaux (Karollus et al., 2023) ont montré que même pour le modèle *Enformer*, la prise en compte des éléments distants reste un défi, car ils ont un poids moindre dans les prédictions du modèle par rapport aux éléments proches.

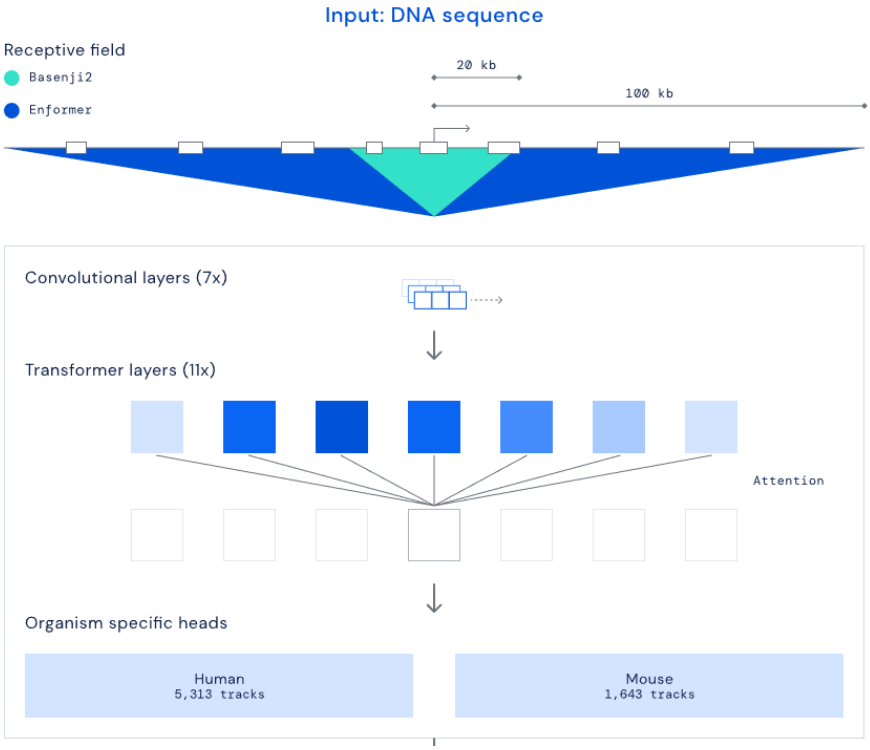


FIGURE 2.15 – **Schéma de l'architecture du modèle enformer.** Ce modèle basé sur les *transformers* permet de prendre jusqu'à 100 000 paires de bases de contexte génomique pour évaluer l'expression d'un gène (AVSEC et al., 2021).

2.3.3 AlphaFold : un transformer pour la conformation 3D des protéines

En 2020, le modèle AlphaFold 2 (JUMPER, 2021b) a remporté la CASP14 (https://predictioncenter.org/casp14/), compétition de référence dans le domaine de la prédiction de structure protéique. Leur modèle a obtenu un score global de 244 (somme de z-scores), contre 90,8 pour le second meilleur modèle (JUMPER, 2021a). Une des caractéristiques principales du modèle AlphaFold est l'usage d'une architecture similaire aux *transformers* (figure 2.16). Ces blocs nommés "*evoformers*" utilisent les mêmes mécanismes d'attention à l'œuvre dans la structure classique des *transformers*. Comme dans le contexte génomique, les interactions 3D qui ont lieu dans les séquences protéiques ne sont pas que locales, ainsi il est nécessaire de pouvoir prendre en compte des interactions longue distance pour obtenir une prédiction de qualité, grâce à ces mécanismes d'attention. L'utilisation de cette architecture a permis le développement d'un outil qui révolutionne le domaine de la biologie structurale rendant possible l'accès à la structure 3D prédite de n'importe quelle protéine du vivant.

Bien que les *transformers* présentent une grande avancée dans le domaine des DNN avec la prise en compte des interactions longue distance, ils présentent un certain nombre de limites qui doivent être considérées avant leur utilisation. Tout d'abord en terme de cout de calcul, les *transformers* par leurs mécanismes d'attention (processus quadratique) sont très couteux en mémoire et en ressources pour le traitement de longues séquences. Par exemple, Alphafold ne prédit que les protéines d'une taille inférieure à 150 acides aminés en raison de la complexité des calculs nécessaires, ce qui exclut un certain nombre de protéines humaines, dont la titine, une

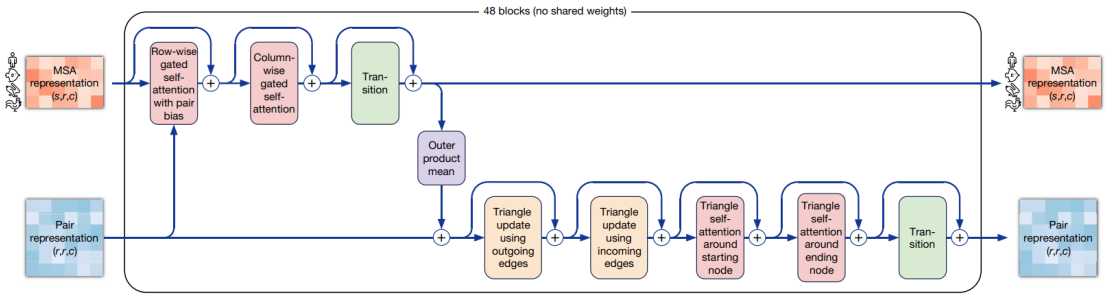


FIGURE 2.16 – **Schéma de l'architecture du modèle AlphaFold.** Ce modèle utilise les mécanismes d'attention des transformers pour prédire la conformation 3D de protéines à partir de leurs séquences et alignements.

protéine essentielle à la structure des muscles. Ensuite, les *transformers* ont besoin d'une quantité de données plus importantes pour leur entrainement que des architectures moins complexes, ce qui peut être un problème dans le domaine biomédical (WILLEMINK et al., 2022). Enfin, alors que pour les CNN il existe des méthodes d'explicabilité *ad-hoc*, les *transformers* étant plus complexes et récents, les techniques d'explicabilité à disposition sont moindre, mais c'est un domaine de recherche actif (KIM et al., 2022; SAHA et al., 2022).

Ainsi l'utilisation des *transformers* est à réserver pour les tâches qu'on ne peut pas résoudre par des architectures plus simples. Une de ces tâches est la compréhension du langage naturel, donc l'architecture *transformer* a permis de révolutionner le traitement grâce aux LLMs que nous présentons dans la prochaine section.

2.3.4 Traitement de langage naturel pré modèle linguistique de grande taille

Le traitement de langage naturel (*Natural Language Processing*, NLP) est une branche de l'IA qui s'intéresse à la compréhension du langage humain de manière automatique. Le NLP est très utile par exemple pour exploiter et extraire de l'information de comptes rendus médicaux en texte libre. Les méthodes traditionnelles de NLP reposent sur le concept d'ontologies, présentées dans le premier chapitre (1.1.3). Cette méthode se base sur une recherche de correspondance exacte des mots d'un texte par rapport à l'ontologie. Cette approche a été implémentée dans divers outils, comme *EMERSE* (HANAUER et al., 2015), pour extraire de l'information des dossiers de santé électroniques de patients. Cependant, elle présente des limites importantes : le besoin d'établir une ontologie exhaustive avec des synonymes, dans la même langue que le texte et que le texte soit exact (sans erreurs d'orthographe ni d'acronymes).

Des méthodes plus flexibles basées sur IA par réseaux de neurones (hors transformers) ont été développées. Ces méthodes reposent sur l'entrainement de réseau de neurones pour la détection et l'extraction d'informations de textes (symptômes, gènes, nom de maladies...). Cette approche a permis le développement, par exemple, de *DeepPhe* (SAVOVA et al., 2017), un outil d'extraction de phénotypes à partir de comptes rendus cliniques ou encore les modèles intégrés dans *SciSpacy* (NEUMANN et al., 2019) capables de détecter et de classer les termes biomédicaux relatifs aux cellules, acides aminés, noms de gènes, noms de maladies, protéines, organismes, tissus, organes et autres. Ces méthodes sont plus robustes que celles basées sur les ontologies, car elles sont résistantes aux erreurs dans le texte et reposent sur une compréhension sémantique des termes et non sur une correspondance exacte de mots. Mais elles présentent tout de même encore des limites. En effet, ces méthodes sont spécifiques à une seule tâche de

reconnaissance et à une langue, elles ne peuvent pas suivre des instructions. S'il n'existe pas de modèles de reconnaissance des termes histologiques ou de modèles entrainés sur des comptes rendus français, il est alors nécessaire de ré-entrainer un modèle à partir d'un nouveau jeu de données.

Les modèles de types *transformers* et plus spécifiquement les LLMs révolutionnent le domaine du NLP en permettant de dépasser les limites des méthodes précédentes. Ces grands modèles dits "Modèles de fondation" (*foundation models*) sont présentés plus en détail dans la section suivante.

2.3.5 La ruée vers l'or des modèles linguistiques de grande taille

Les années 2022-2023 représentent des années clés dans l'histoire de l'analyse et la compréhension de texte (NLP) avec le développement et la mise à disposition de plusieurs modèles de fondation généraux performants et accessibles tel que *GPT-3.5-turbo* (souvent nommé à tort *ChatGPT)* ou *LLaMA* (Touvron et al., 2023). Ces LLMs représentent une véritable révolution dans le traitement des données textuelles non structurées, car ils sont capables de suivre des instructions précises et extrêmement variées, dans de multiples langues et de manière robuste face aux erreurs dans le texte. Dans un long papier de 155 pages nommé "*Sparks of Artificial General Intelligence : Early experiments with GPT-4*" publié en avril 2023 par une partie de l'équipe ayant travaillé sur le modèle GPT-4 d'OpenAI (Bubeck et al., 2023), les auteurs dressent un portrait des capacités des LLMs et de leur possible impact sociétal.

Pour ne citer que quelques exemples des capacités des LLMs, ces modèles sont capables de résumer des articles scientifiques en une série de points clés, de générer du code informatique et d'en corriger les erreurs, d'extraire des informations spécifiques d'un texte, de répondre à des questions de logique, d'expliquer des blagues, d'écrire des histoires créatives, d'analyser des images et d'interagir avec le monde extérieur par des requêtes internet. Ces modèles ont donc des capacités extrêmement vastes avec des performances proches de l'Homme. Les modèles LLMs sont très divers et en constante évolution. La figure 2.17 (J. Yang et al., 2023) retrace les principaux modèles développés entre 2018 et 2023 où l'on observe une nette évolution entre 2022 et 2023 avec une multitude de modèles ouverts ou fermés développés.

Cette évolution et accélération de la recherche que j'ai choisi de qualifier de "ruée vers l'or" traduit un net engouement pour ces modèles et des tâches qui peuvent être accomplies. Cette émulation de la recherche en IA autour des LLMs a mené au développement de modèles de toutes tailles et de tous types ainsi qu'au développement d'infrastructures et de techniques pour l'hébergement, l'optimisation et l'utilisation de ces nouveaux modèles. Le développement de tels modèles a été rendu possible par l'utilisation de deux innovations en plus des *transformers* : l'apprentissage auto-supervisé (*self-supervised learning, SSL*) et l'apprentissage par renforcement avec retour humain (*Reinforcement Learning from Human Feedback, RLHF*) que nous allons brièvement présenter ci-dessous.

2.3.5.1 Apprentissage auto-supervisé

En mars 2021, Yann Lecun a écrit un billet de blog qualifiant l'apprentissage auto-supervisé comme la matière noire de l'intelligence *(Self-supervised learning : The dark matter of intelligence*, LeCun et Misra, 2021). Dans ce billet, il explique que l'apprentissage auto-supervisé est l'une des voies les plus prometteuses pour construire des connaissances de fond dans les modèles IA.

L'idée dernière le SSL part d'un constat : l'ensemble des modèles d'IA nécessitent un jeu de données bien annotées pour leur entrainement. Or, il est facile d'acquérir un grand nombre de données non annotées, le goulot d'étranglement se situe dans l'annotation et la labellisation

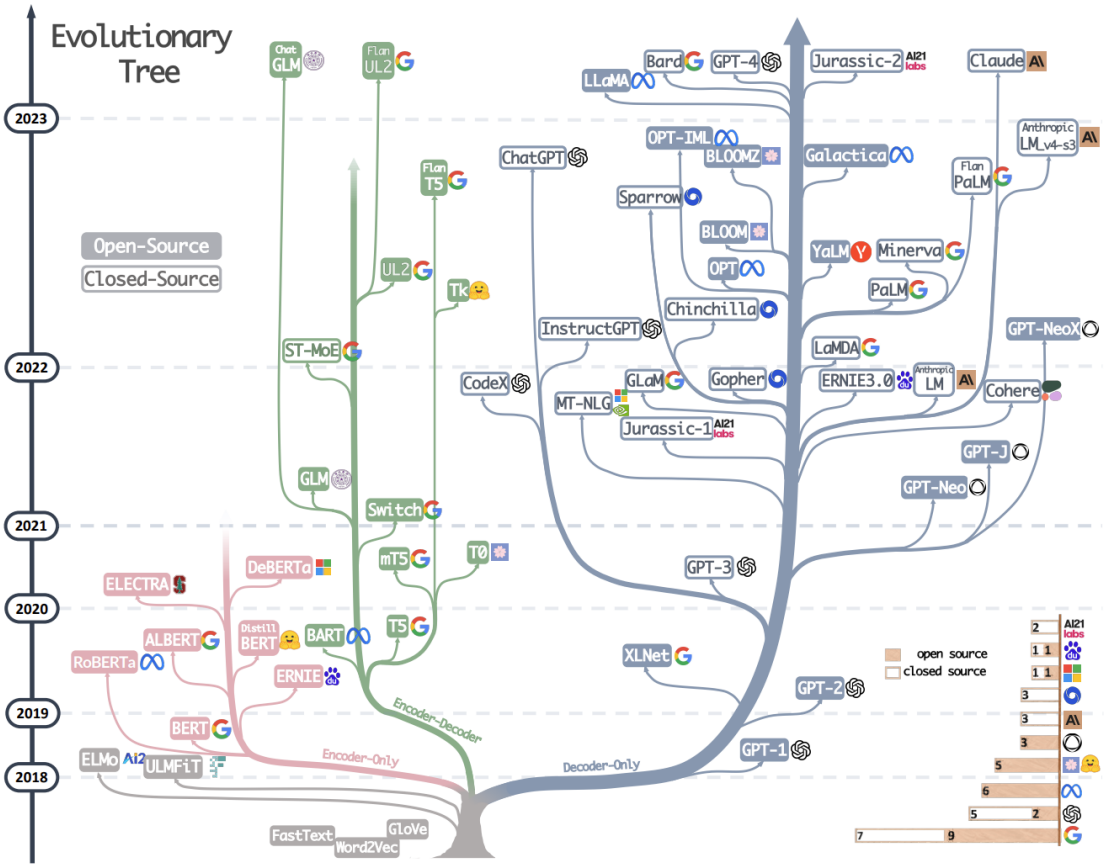


FIGURE 2.17 – **Arbre du développement LLMs de 2018 à 2023**. On observe à partir de 2022 une forte croissance du nombre de modèles développés pour l'analyse de texte basés sur les architectures de *transformers*. (J. YANG et al., 2023)

des jeux de données. Par exemple, il est facile d'acquérir un grand nombre de pages internet et d'en extraire le texte. Comment utiliser ces données non annotées pour entrainer un modèle de langage ? Est-il possible de définir une méthode d'apprentissage à partir des données non annotées ? C'est précisément l'intérêt du SSL.

Le fonctionnement du SSL est relativement simple. Dans une phase de pré-entrainement, le modèle va apprendre à recréer les données qui lui sont fournies. Par exemple, si l'on veut créer un modèle de langage, on peut lors de ce pré-entrainement lui fournir une phrase telle que :
"Les modèles IA sont des systèmes complexes capables de réaliser des prédictions".
Pour le pré-entrainement du modèle par SSL on va cacher certains mots de la phrase et entrainer le modèle à prédire le mot attendu. Ainsi on obtient la phrase :
"Les modèles IA sont des systèmes [?] capables de réaliser des [?]".
Et pour chaque point d'interrogation, le modèle va être entrainé à prédire les mots "complexes" et "prédictions" avec une grande probabilité. Par cette méthode, il est possible d'utiliser des données non annotées pour réaliser un pré-entrainement qui permet au modèle une meilleure compréhension de la structure d'un texte et de la relation entre les mots. Après ce pré-entrainement, l'entrainement classique du modèle peut avoir lieu, avec un nombre de données annotées plus limité que sans le pré-entrainement par SSL.

Le concept de SSL n'est pas limité au texte, le principe est similaire avec des images. Par exemple, il est possible de diviser une image en 16 zones carrées, de masquer une de ces zones et d'entrainer le modèle à régénérer la zone en guise de pré-entrainement. Pour cela, le jeu de données ImageNet (Deng et al., 2009) est communément utilisé. Ce jeu de données est composé de plus de 14 millions d'images réparties sur 1000 classes différentes et variées (animaux, objets, fruits, bâtiments). Il est typiquement utilisé comme méthode d'évaluation des performances des nouvelles architectures de réseaux de neurones. Il a été montré qu'un pré-entrainement sur ImageNet par SSL permet d'obtenir de meilleures performances de classification avec une quantité de données moindre (Goyal et al., 2021)

2.3.5.2 Apprentissage par renforcement avec retour humain

L'apprentissage par renforcement avec retour humain est une méthode qui a permis l'amélioration des performances des LLMs (Stiennon et al., 2020; Ziegler et al., 2020). Cette méthode permet non seulement de réduire les biais dans le modèle grâce à une intervention humaine (biais raciaux, de genre, religieux, Ganguli et al., 2022), mais aussi de créer un modèle capable de suivre des instructions (Ouyang et al., 2022).

Après le pré-entrainement par SSL et l'entrainement classique du modèle avec des données annotées, les sorties du modèle vont être affinés par apprentissage par renforcement. Plusieurs exemples de générations pour une instruction vont être présentés à un humain qui va devoir classer les réponses dans un ordre de préférence. Le modèle va ainsi obtenir une pénalité ou une récompense pour chaque réponse, ce qui permet d'affiner ses futures réponses. Ainsi le retour humain est intégré à l'entrainement du modèle par ce mécanisme d'apprentissage par renforcement pour rentre le modèle moins biaisé et qualitatif.

2.3.6 Les modèles génératifs et modèles d'*embedding*

Les LLMs se déclinent sous deux formes principales avec des utilisations différentes : les modèles génératifs et les modèles d'*embedding*. La figure 2.18 présente les deux types de modèles.

Les modèles de type *embedding* consistent à transformer un texte d'entrée en vecteur numérique de grande taille. L'idée derrière cette transformation est que deux phrases ayant un sens similaire (sens sémantique) ont une représentation numérique similaire. Transformer le texte en

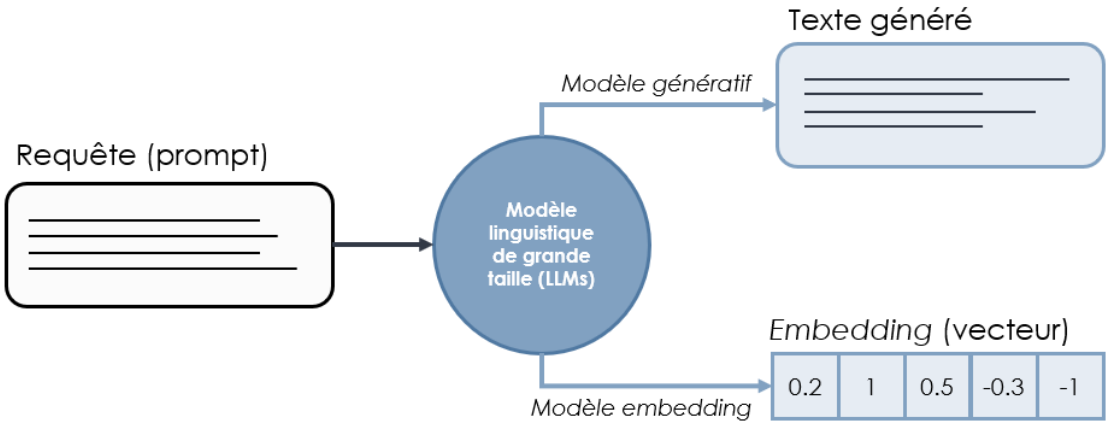


FIGURE 2.18 – **Schéma des deux types de LLMs principaux**. D'un côté, il y a les modèles génératifs qui sont des modèles de complétions, produisant du texte à partir de texte. De l'autre, les modèles d'*embedding* transforment le texte d'entrée en vecteur numérique de grande taille représentant le sens sémantique du texte d'entrée.

vecteur numérique permet de réaliser des opérations, telles que la recherche de similarité entre deux phrases, mais aussi le *clustering*, la visualisation graphique ou encore l'entrainement de modèles de classification. Le concept d'*embedding* est donc un réel couteau suisse pour l'analyse et la recherche de similarité entre des textes et il est en général multilingue : deux phrases identiques, mais dans des langues différentes vont avoir une représentation numérique similaire. La figure 2.19 (LUIS SERRANO, 2023) présente une visualisation de l'*embedding* de 9 phrases en deux dimensions. On observe sur cette figure que les phrases traitant d'un sujet similaire sont regroupées en sous-groupe dans l'espace. On observe un groupe qui concerne les chiens, un groupe concernant le football et un groupe concernant une question de type "comment vas-tu". Cet exemple est simplifié à des fins de représentation, car les phrases ne sont encodées qu'en deux dimensions. En réalité pour capturer des nuances complexes dans les phrases, les modèles d'*embedding* encodent les phrases en plusieurs centaines voire en milliers de dimensions.

Les modèles génératifs sont des modèles de complétion. À partir d'un texte en entrée, ils génèrent un texte de sortie. Ces modèles sont utiles pour des tâches créatives comme la génération d'histoires ou de code, mais aussi pour suivre des instructions précises telle l'extraction d'informations d'un texte comme la récapitulation d'un article scientifique. Le texte en entrée contenant les instructions pour le modèle est nommé *prompt* (figure 2.20). Le défi dans l'utilisation des modèles génératifs réside dans la création d'un *prompt* adapté aux résultats que l'on souhaite obtenir. Le processus de création et d'affinage d'un *prompt* pour une tâche se nomme "*prompt engineering*". Classiquement, un *prompt* possède une structure de base en quatre éléments : (i) la description précise de la tâche (ii) un exemple de réalisation de la tâche et de résultat souhaité(iii) le texte que l'on souhaite analyser (iv) un indicateur de sortie indiquant au modèle qu'on attend une réponse. L'avantage des modèles génératifs réside dans le fait qu'aucun apprentissage supplémentaire n'est nécessaire, le point central étant l'établissement d'un *prompt* optimal.

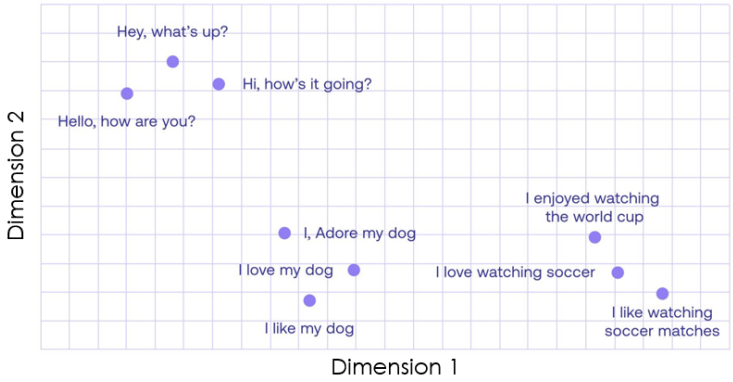


FIGURE 2.19 – **Visualisation d'embedding de 9 phrases en 2 dimensions**. Dans cette représentation en deux dimensions, on observe que les phrases qui concernent des sujets similaires se regroupent dans l' espace en clusters, indiquant que leurs représentations numériques (par *embedding*) sont similaires. (LUIS SERRANO, 2023)

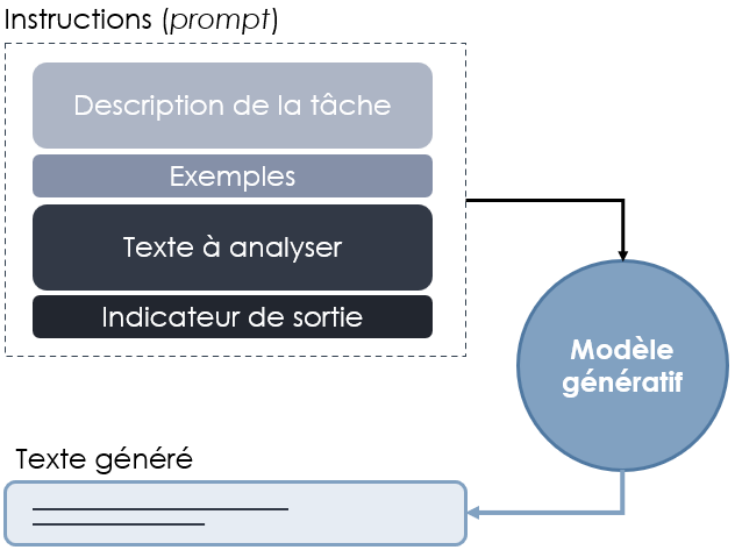


FIGURE 2.20 – **Schéma du concept de *prompt* pour les LLMs génératifs**. Un *prompt* typique est constitué de quatre parties : (i) une description de la tâche, (ii) un exemple de réalisation de la tâche, (iii) le texte à analyser en entrée, et (iv) un indicateur de sortie.

2.3.7 Taille des modèles et défis d'utilisation

Les LLMs sont très versatiles et peuvent accomplir de nombreuses tâches, mais ils restent une technologie très récente en constante évolution et encore peu optimisée. Bien que les modèles d'*embedding* sont très facilement utilisables sur presque n'importe quel matériel informatique, les modèles génératifs quant à eux sont en général d'une taille extrême rendant leur hébergement très compliqué. Comme déjà abordé dans la section "Nombre de paramètres et ressources informatiques" ci-dessus, les LLMs peuvent avoir une taille allant de 10 à plusieurs centaines de milliards de paramètres, rendant difficile leur adoption et leur utilisation.

Pour rendre ces modèles accessibles, plusieurs stratégies existent. La première est la mutualisation de l'hébergement, où une entité (une entreprise, un laboratoire), va héberger sur du matériel informatique spécialisé un LLMs qui va être disponible *via* des requêtes internet (*Application Programming Interface*, API). Cette solution pose des problèmes de confidentialité des données.

La deuxième stratégie est d'optimiser les modèles de grande taille. Une première méthode d'optimisation utilisée est nommée la quantisation (Dettmers et al., 2022; Dettmers et Zettlemoyer, 2023) qui consiste à réduire la complexité informatique des poids des modèles. Les poids des modèles sont représentés par des nombres décimaux sur 16 bits informatiques (65 536 valeurs possibles). Il est possible de réduire cette complexité en représentant les poids du modèle sur seulement 4 bits (16 valeurs possibles) avec une perte minimale en précision, ce qui permet de réduire d'un facteur 4 la taille du modèle et les couts d'hébergement.

La troisième stratégie est d'entrainer des modèles de plus petite taille à partir des sorties de modèles de grande taille, afin de les "imiter". Ainsi, il a été possible d'entrainer de petits modèles avec seulement 7 milliards de paramètres (contre 175 milliards pour GPT-3) à partir de 52 000 exemples de générations de textes de grands modèles (Peng et al., 2023). Cependant, de récents travaux (Gudibande et al., 2023) ont montré que cette approche bien qu'efficace en apparence, n'est pas capable d'imiter l'ensemble des capacités des modèles de très grande taille.

Pour conclure ce chapitre, il est clair que les méthodes présentées permettent d'explorer de façon rétrospective et multimodale, de nombreuses données biomédicales acquises sur des patients, qu'elles soient textuelles, d'imagerie ou de séquences. Ces méthodes, bien que complexes à mettre en place et à utiliser, permettent de combler les limites du ML présenté dans le chapitre précédent et ouvrent la porte à l'exploitation de données sans nécessiter un travail d'annotation et de structuration intensif. La stratégie de couplage de réseaux de neurones comme extracteurs d'informations de données non structurées avec des algorithmes de ML transparents et explicables apparait prometteuse pour l'exploitation des données biomédicales non structurées. Dans le prochain chapitre, nous présenterons un cas d'application de ces nouvelles méthodes à travers les myopathies congénitales, une famille de maladies génétiques rares. Nous présenterons leur diagnostic et les données associées à celui-ci.

Chapitre 3

L'exemple des myopathies congénitales et la difficulté du diagnostic

Dans cette thèse, nous avons cherché à développer des méthodes IA pour exploiter les données biomédicales. Nous nous sommes concentrés sur les données de patients atteints de myopathies congénitales : une famille de maladies des muscles rares et génétiques, dont le diagnostic est complexe. Dans ce chapitre, nous allons présenter le contexte biologique de ces maladies avec une présentation de la structure du muscle, de la classification des myopathies congénitales et de leur diagnostic.

3.1 Le muscle, un organe particulier assurant des fonctions diverses

Le muscle est un organe particulier tant par sa structure que par son abondance dans le corps humain. Pour un adulte, les muscles représentent près de 40% de la masse corporelle totale (Janssen et al., 2000). Les muscles assurent une variété de fonctions, dont le mouvement et la posture, mais aussi la thermogénèse et l'équilibre métabolique.

Dans le muscle, divers processus sont nécessaires pour permettre d'assurer ses fonctions. Le muscle intègre des processus neurologiques pour le transfert de signaux induisant la contraction. Cette contraction est possible grâce à une structure particulière du muscle en fibres que nous présenterons en détail dans la section 3.1.2. De plus, cette contraction requiert de l'énergie qui est fournie par des processus métaboliques. Enfin, le muscle, en raison des contraintes mécaniques subies lors des mouvements, est un organe dynamique dans lequel des processus de régénération musculaire sont à l'œuvre pour assurer sa plasticité.

3.1.1 Types de muscles

Il existe dans le corps humain trois types de muscles avec des structures différentes selon la fonction qu'ils doivent assurer (figure 3.1, Gómez Oca, 2021) : le muscle lisse, le muscle cardiaque et le muscle strié squelettique.

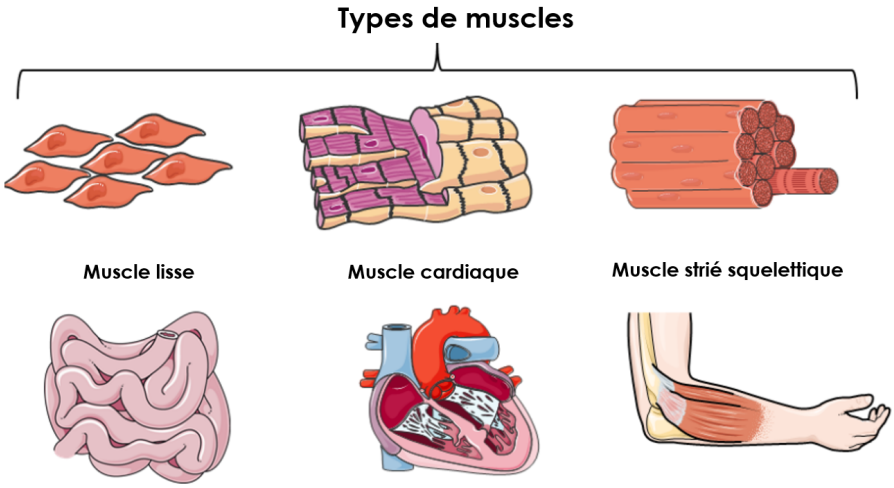


FIGURE 3.1 – **Schéma des trois types de muscles**. Les muscles lisses se contractent de manière involontaire (vaisseaux sanguins, appareil digestif). Les muscles striés squelettiques se contractent de manière volontaire (mouvement du corps). Les muscles cardiaques sont aussi striés, mais se retrouvent uniquement dans le cœur et se contractent de manière involontaire et en rythme. (traduit de GÓMEZ OCA, 2021)

3.1.1.1 Muscle lisse

Le muscle lisse, aussi nommé non strié tire son nom de l'absence de striations lors de l'observation au microscope. Les fibres ne possèdent qu'un seul noyau central. Ce type de muscles est présent dans la paroi de nombreux organes, tels que les vaisseaux sanguins, l'appareil digestif, l'appareil respiratoire, l'appareil urinaire et la paroi des viscères. La particularité des muscles lisses est qu'ils se contractent de manière involontaire, sans contrôle conscient. La contraction de ces muscles est régie par le système nerveux neurovégétatif (système autonome).

3.1.1.2 Muscle strié cardiaque

Le muscle strié cardiaque est un muscle qui se retrouve exclusivement dans le cœur. Il partage des caractéristiques communes à la foi au muscle lisse et au muscle strié squelettique. Ce muscle se contracte de manière involontaire, et les fibres ne possèdent qu'un seul noyau central à l'instar des muscles lisses. Cependant, le muscle cardiaque présente des striations similaires au muscle strié squelettique. La caractéristique unique du muscle strié cardiaque est qu'il fonctionne en continu et se contracte en rythme de façon coordonnée. Ainsi ce muscle est très dépendant du métabolisme oxydatif.

3.1.1.3 Muscle strié squelettique

Enfin, les muscles striés squelettiques, aussi nommés muscles volontaires, sont les muscles mobilisés lors des mouvements conscients et volontaires, lorsque l'on porte un objet ou que l'on fait du sport par exemple. Au microscope, ces muscles se démarquent par la présence de stries transversales et longitudinales. Les cellules composant les fibres musculaires du muscle strié squelettique ont la particularité d'avoir fusionné, formant ainsi un "syncitium vrai". Cette fusion donne donc aux fibres musculaires les caractéristiques des cellules géantes (entre 1 et 5

cm de long et 10 à 100µm de diamètre). Ainsi, ces cellules sont multinucléées avec des noyaux périphériques.

Comme leur nom l'indique, ces muscles sont reliés aux os par l'intermédiaire du tendon, leur contraction permet donc le mouvement des os et donc du corps. Cette contraction est engagée sous le contrôle du système nerveux somatique. En fonction de l'intensité et de la durée de la contraction, les muscles striés squelettiques peuvent fonctionner de manière aérobie grâce à leur vascularisation importante ou de façon anaérobie.

Dans le cadre des myopathies congénitales, les muscles striés squelettiques sont affectés et ne sont plus capables d'assurer leur fonction normale en raison d'altérations structurelles, entrainant un déficit de tonus musculaire et de force.

3.1.2 Structure du muscle strié squelettique

Pour comprendre comment les altérations de la structure du muscle observées dans les myopathies congénitales peuvent mener à son dysfonctionnement, il est important de comprendre comment le muscle est structuré et comment il peut se contracter.

L'organisation du muscle peut être décrite comme celle d'une corde d'escalade. Une corde d'escalade semble être composée d'un seul élément fort et résistant. Mais si l'on y regarde de plus près, une corde d'escalade est constituée d'une gaine qui entoure l'âme de la corde. Cette âme est composée de plusieurs gros filaments, eux-mêmes composés de plusieurs filaments de plus en plus fins, torsadés ensemble.

Ainsi par analogie, le muscle entier est comme la corde d'escalade, sa structure est décrite dans la figure 3.2 (Burr et Allen, 2019). Le muscle entier, entouré par sa gaine nommée *epimysium* et *perimysium*, est constitué de plusieurs faisceaux musculaires qui le composent. Chacun de ces faisceaux ou fascicules est composé de filaments encore plus fins nommés fibre musculaire. Ces fascicules et fibres musculaires sont encore observables au microscope optique. Enfin, chacune de ces fibres musculaires est en fait un regroupement de plus petits filaments nommés myofibrilles qui sont des filaments composés d'une multitude de myofilaments capables de se contracter. Les myofibrilles et myofilaments ne sont observables qu'en microscopie électronique.

À une échelle encore plus petite, les myofibrilles sont organisées en plusieurs sarcomères, présentés en figure 3.3 (Burr et Allen, 2019). Le sarcomère est l'unité contractile de base des myofibrilles. Il est constitué de trois systèmes de filaments protéiques : (i) un filament épais de myosine (ii) un filament mince d'actine inséré sur le disque Z et (iii) un filament élastique enchâssé sur le filament de myosine composé de la titine insérée sur le disque Z aussi.

Lors d'une contraction musculaire, les filaments de myosine et d'actine dans la bande A glissent les uns sur les autres, réduisant ainsi la zone H, le muscle est contracté, le sarcomère est raccourci. On comprend alors qu'un dysfonctionnement dans l'une de trois protéines essentielles à ce mouvement de contraction altère la structure du sarcomère et donc la capacité contractile du muscle. Les gènes TTN, ACTA1 et MYH2/MYH7 (codant respectivement pour la titine, actine et myosine) sont des gènes typiquement responsables de myopathies lorsqu'ils sont mutés. Cependant les gènes impliqués dans le sarcomère ne sont pas les seules pouvant provoquer un dysfonctionnement musculaire, nous verrons l'ensemble des gènes responsables des myopathies congénitales dans une section suivante.

3.1.3 Types de fibres musculaires

Les fibres musculaires peuvent être classées en trois types de fibres : les fibres de type 1, les fibres de type 2A et 2B qui ont des caractéristiques structurelles et fonctionnelles différentes. Cette classification repose sur le profil d'expression de la chaine lourde de la myosine. Il existe

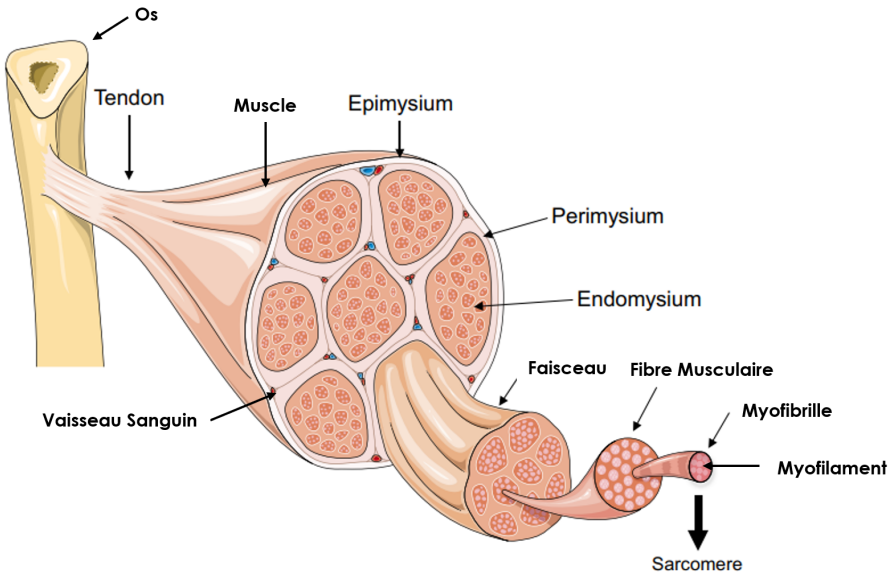


FIGURE 3.2 – **Schéma de la structure du muscle strié squelettique**. Le muscle lié à l'os par le tendon est structuré en plusieurs faisceaux entourés par le *perimysium* et l'*epimysium*. Chacun de ces faisceaux est composé de fibres musculaires elles-mêmes composées de myofibrilles et de myofilaments. Dans chacun de ces myofilaments, on retrouve plusieurs sarcomères. (modifiée de BURR et ALLEN, 2019).

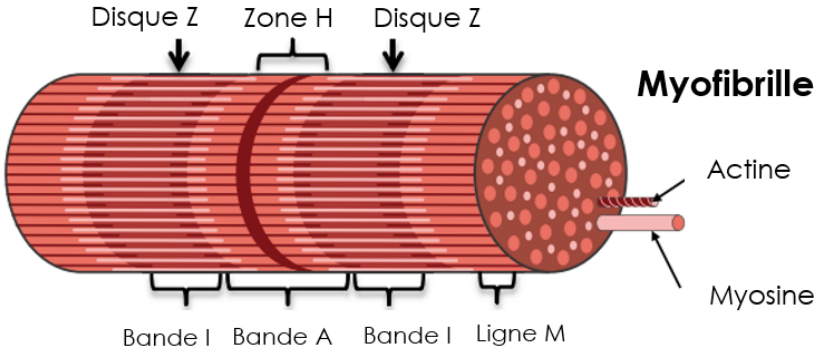


FIGURE 3.3 – **Schéma de la structure du sarcomère**. Le sarcomère est composé de trois filaments protéiques : un filament épais de myosine, un filament mince d'actine inséré sur le disque Z et un filament élastique de titine et myosine. Par glissement de la myosine sur l'actine dans la bande A, la contraction est rendue possible, raccourcissant la zone H. (modifié de BURR et ALLEN, 2019).

plusieurs isoformes de la myosine (MYH1, MYH2 et MYH7 Tajsharghi et Oldfors, 2013), permettant chacun une contraction plus ou moins rapide et résistante à la fatigue. Le tableau 3.1 répertorie les différences principales entre les fibres de type 1, 2B et 2A.

Caractéristique	Fibre Type 1	Fibre Type 2B	Fibre Type 2A
Couleur	Rouge	Pâle	Pâle
Vitesse de Contraction	Lente	Rapide	Rapide
Voie métabolique	Aérobie	Anaérobie	Aérobie et anaérobie
Réserve en oxygène	Importante	Faible	Moyenne
Réserve en glycogène	Faible	Importante	Moyenne
Fatigabilité	Faible	Importante	Moyenne

Tableau 3.1 – **Tableau de comparaison des fibres de type 1, 2B et 2A.** Les fibres de type 1 sont des fibres à contraction lente, endurantes et utilisant le métabolisme aérobie. Les fibres types 2B sont des fibres à contraction rapide, avec faible endurance et à voie métabolique anaérobie. Les fibres de type 2A sont à mi-chemin entre les fibres de type 1 et 2B en termes d'endurance et de vitesse de contraction.

Les fibres de type 1, nommées fibres rouges, sont des fibres à contraction lente. Elles ont en général un plus petit diamètre et sont plus vascularisées. Ce sont des fibres spécialisées dans l'aérobie et très résistantes à la fatigue. Elles sont efficaces dans l'utilisation de l'oxygène pour générer de l'adénosine triphosphate (ATP) qui est la source d'énergie permettant la contraction musculaire. Leur couleur rouge provient de la présence de myoglobine, la protéine stockant l'oxygène dans le muscle. Ces fibres sont utilisées pour le maintien de la posture et sont mobilisées dans les activités d'endurance à faible intensité.

Les fibres de type 2B sont des fibres à contraction rapide. Ces fibres sont aussi nommées fibres blanches, elles sont spécialisées dans les mouvements rapides et explosifs. Ces fibres utilisent des voies anaérobies pour générer de l'ATP et se fatiguent beaucoup plus rapidement. La voie anaérobie ne repose pas sur l'oxygène, mais l'utilisation du glycogène. Ces fibres possèdent donc des réserves de glycogène plus importantes que les fibres de type 1. Ces fibres sont mobilisées dans le cadre d'activité intense telle que des sprints.

Enfin, les fibres de type 2A représentent un intermédiaire entre les fibres de type 1 et 2B. Ces fibres peuvent générer de l'énergie (ATP) à la fois par la voie aérobie et anaérobie. À l'instar des fibres 2B, elles peuvent générer des contractions puissantes, mais se fatiguent rapidement. Ces fibres sont mobilisées lors des exercices à intensité moyenne demandant de l'endurance telle que la natation.

La balance entre fibres type 1 et 2A/2B dans un muscle est un marqueur important des myopathies congénitales. Souvent dans le cadre de myopathies, une prédominance des fibres de type 1 va se manifester dans le muscle. En microscopie, la visualisation de ces fibres se réalise par des méthodes histochimiques telles que la coloration ATPase qui colore différentiellement les fibres de type 1 et fibres de type 2.

3.1.4 Classification des atteintes neuromusculaires

La variété des processus impliqués dans le fonctionnement normal du muscle strié squelettique ouvre la porte de nombreux dysfonctionnements pouvant amener à une défaillance du muscle. Ces défaillances peuvent provoquer des maladies aux manifestations variées que l'on nomme les atteintes neuromusculaires (*neuromuscular diseases*, NMD). Les NMD ont une prévalence d'environ 3,7 à 4,9 pour 10 000 (Lace et al., 2022), ce qui les classe parmi les maladies

rares (prévalence inférieure à 5 pour 10 000). Les NMD héréditaires affectant les muscles striés squelettiques peuvent être classifiées parmi quatre grandes catégories présentées dans le tableau 3.2 (Benarroch et al., 2023; Lornage, 2019). Les dystrophies musculaires sont caractérisées principalement par une perte progressive de la force et de la masse musculaire. Les patients atteints de myopathies métaboliques présentent une forte intolérance à l'exercice et des épisodes de fatigue. Les myopathies mitochondriales sont aussi caractérisées par une faiblesse musculaire et une intolérance à l'exercice, mais en plus par des problèmes cardiaques, auditifs et des crises d'épilepsie.

Enfin, les myopathies congénitales, qui sont le sujet principal de notre cas d'application, sont des maladies avec un départ précoce de la faiblesse musculaire (souvent dès la naissance) et dont la progression est lente. De plus, les patients atteints de myopathies congénitales présentent souvent des caractéristiques faciales dysmorphiques (visage, bouche). Dans la prochaine section, nous allons voir en détail la classification et la prévalence des myopathies congénitales ainsi que leur diagnostic.

Dystrophies musculaires	**Myopathies métaboliques**
— Faiblesse musculaire progressive — Perte de masse musculaire	— Intolérance à l'exercice — Épisodes de fatigue — Myalgie
Myopathies mitochondriales	**Myopathies congénitales**
— Faiblesse musculaire — Intolérance à l'exercice — Implication cardiaque — Perte auditive — Crises d'épilepsie	— Faiblesse musculaire à départ précoce — Progression lente de la maladie — Caractéristiques faciales dysmorphiques (visage allongé et palais vouté)

Tableau 3.2 – **Tableau des différentes atteintes neuromusculaires héréditaires et leurs caractéristiques principales**. Il y a quatre grandes classes d'atteintes neuromusculaires héréditaires affectant le muscle strié squelettique. Dans cette thèse, nous nous sommes concentrés sur les myopathies congénitales. (Lornage, 2019)

3.2 Les myopathies congénitales

3.2.1 Description générale

Les myopathies congénitales (myopathies congénitales (MC)) sont une famille de maladies génétiques rares qui affectent les muscles en général dès la naissance, mais les premiers symptômes peuvent n'apparaitre qu'à l'adolescence ou à l'âge adulte. Les MC se caractérisent principalement par la présence d'anomalies histopathologiques dans la biopsie musculaire indiquant une anomalie de la structure musculaire et de ses capacités contractiles.

Les myopathies congénitales sont des maladies génétiques rares, dont la prévalence d'environ 1,5 pour 100 000 dans la population générale et 2,73 pour 100 000 dans la population pédiatrique (Huang et al., 2021). Elles sont donc une prévalence inférieure à 50 pour 100 000, seuil pour considérer une maladie comme rare. La *Muscle Gene Table* (https://www.musclegenetable.fr/, Benarroch et al., 2023) référence l'ensemble des gènes responsables et des classes de maladie

considérées comme des NMD. Sur les 658 gènes référencés, il y a 47 gènes identifiés comme pouvant être responsables de MC et cette liste évolue chaque année avec l'identification de nouveaux gènes, ce qui rend le diagnostic génétique complexe.

3.2.2 Approches curatives des myopathies congénitales

Actuellement, il n'existe aucun traitement curatif autorisé sur le marché contre les MC. Des approches curatives sont en cours de développement soit au stade préclinique soit au stade d'étude clinique (Gineste et Laporte, 2023 ; Guan et al., 2016). Parmi les approches explorées, on retrouve des approches de thérapies géniques, qui consistent à essayer de rétablir la fonction du gène défaillant en introduisant une copie fonctionnelle du gène dans l'organisme. Cette approche est en cours d'étude pour les MC liées à une défaillance des gènes ACTA1, MTM1 et DNM2, BIN1, RYR1 et STIM1. Cependant, les thérapies géniques sont très couteuses et complexes à développer et mettre en place. Une seconde approche utilisée est l'utilisation de molécules pharmacologiques pour pallier la fonction défaillante. Ces molécules sont en études précliniques dans de nombreuses MC, telles que celle causée par les gènes cités précédemment, ainsi que TPM2, TMP3, NEB et SEPN1.

Cependant, que ce soit pour la thérapie génique ou l'utilisation de molécules pharmacologiques, les développements restent difficiles, notamment dans le cadre du passage de l'expérimentation animale à l'Homme. La startup strasbourgeoise Dynacure a développé en 2019 une petite molécule, un oligonucléotide antisens, nommée DYN101 pour le traitement des myopathies congénitales centronucléaires liées aux gènes DNM2 et MTM1. Le développement de ce traitement a dû être arrêté suite à une toxicité hépatique trop importante chez l'Homme lors de la seconde phase de l'essai clinique. Cette toxicité n'avait pas été observée dans le modèle murin utilisé.

3.2.3 Classification et prévalence

Les myopathies congénitales sont un groupe de maladies hétérogènes dont le principal critère de classification est la biopsie musculaire. Les myopathies congénitales peuvent être classées en trois sous-types principaux : (i) myopathies à *cores* (COM) (ii) myopathies à Némaline (NM) (iii) myopathies centro-nucléaires (CNM) et deux sous-types supplémentaires : (iv) myopathies à disproportion congénitale des fibres (CFTD) et (v) myopathie de stockage de myosine (*myosin storage myopathies*, MSM) (Cassandrini et al., 2017 ; Claeys, 2020 ; North et al., 2014). Dans cette sous-section, nous allons présenter les caractéristiques histologiques et cliniques principales de chaque sous-type ainsi que les gènes impliqués.

3.2.3.1 Myopathies à cores (COM)

Les myopathies à cores (COM) peuvent se diviser en deux sous-groupes supplémentaires : les myopathies à core unique et central (CCD) et les myopathies à multi-minicores (MmD). Les COM sont largement associées à différentes mutations dans le gène RYR1 avec des formes dominantes provoquant principalement le sous-type CCD et récessives provoquant principalement le sous-type MmD. Les myopathies à core de type MmD possèdent un terrain génétique plus varié, des mutations dans SEPN1 et MYH7 pouvant aussi être responsables de la maladie (Cassandrini et al., 2017). Ce sous-type de myopathies possède une prévalence d'environ 0,37 pour 100 000 (Huang et al., 2021) en faisant la MC la plus commune.

Au niveau histopathologique, les COM se caractérisent principalement par la présence de cores (présentés en figure 3.4), des zones avec une activité oxydative très réduite. Ces zones sont soit uniques, centrales et de grande taille, soit de petites tailles et multiples dans une même

fibre musculaire. De plus, on retrouve moins fréquemment la présence de noyaux centralisés, une disproportion dans le ratio de fibres type 1 et type 2 et la présence d'infiltration du tissu conjonctif (Jungbluth et al., 2018). Au niveau clinique, en fonction du gène impliqué, on va retrouver très fréquemment des atteintes des muscles extra-oculaires et des atteintes bulbaires (RYR1 récessif), ainsi que des atteintes respiratoires et cardiaques (SEPN1). Les symptômes des COM peuvent apparaitre à la naissance, durant l'enfance où à l'âge adulte en fonction du gène impliqué.

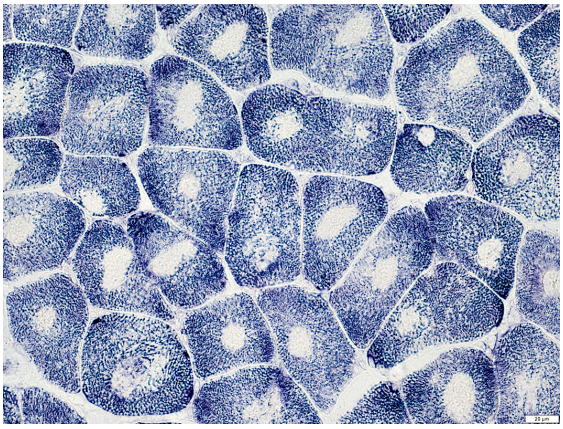


Figure 3.4 – **Biopsie de muscle présentant des cores**. Ces zones à faible activité oxydative sont caractéristiques des myopathies à core à la coloration NADH (Alan Pestronk, 2022)

3.2.3.2 Myopathies à némaline (NM)

Les myopathies à némaline (NM) sont associées à plus d'une dizaine de gènes, dont le plus communément responsable est NEB (Jungbluth et al., 2018). Ce sous-type de myopathie présente une prévalence de 0,20 pour 100 000 (Huang et al., 2021). Au niveau histopathologique, cette myopathie se caractérise principalement par la présence d'inclusions ressemblant à des bâtonnets (illustré en figure 3.5). Il peut aussi y avoir la présence de cores similaires aux COM. Au niveau clinique, les patients atteints de NM présentent des problèmes respiratoires, des contractures et des atteintes bulbaires (Jungbluth et al., 2018), mais pas d'atteinte cardiaque. Les symptômes apparaissent en général dès la naissance et parfois à l'enfance, mais pas à l'âge adulte (Jungbluth et al., 2018).

Les NM possèdent un niveau de sous-typage supplémentaire avec deux classes : (i) les myopathies à "cap", une forme très rare de NM avec 20 patients décrits entre 1981 et 2017 et (ii) les myopathies nommées "*zebra body*" où le muscle possède une apparence zébrée, qui est une forme de MC bénigne avec moins de 10 patients décrits (Cassandrini et al., 2017).

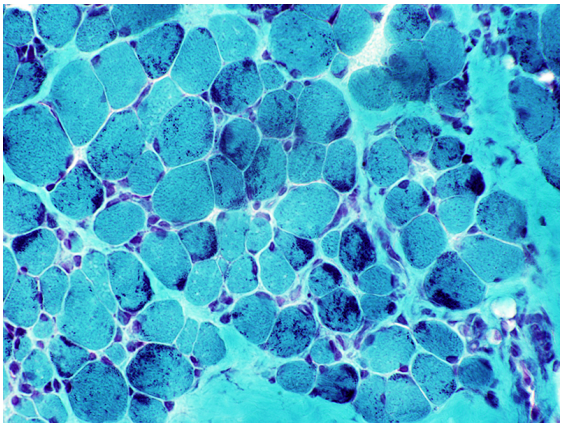


FIGURE 3.5 – **Biopsie de muscle présentant des bâtonnets**. Ces inclusions sombres sont caractéristiques des myopathies à némaline à la coloration trichrome de Gomori (ALAN PESTRONK, 2022).

3.2.3.3 Myopathies centronucléaires (CNM)

Les myopathies centronucléaires (CNM), aussi nommées myopathies myotubulaires, sont des myopathies congénitales plus rares avec une prévalence d'environ 0,08 pour 100 000 (HUANG et al., 2021). Ce groupe de myopathies peut-être divisé en deux sous-groupes : les CNM liées au chromosome X (XLMTM) dû au gène MTM1 et non liées à l'X (DNM2, RYR1, BIN1, TTN...) (NORTH et al., 2014). Dans une fibre musculaire saine, les noyaux sont positionnés en périphérie des fibres. Ce groupe de MC se caractérise au niveau histopathologique par la présence de noyaux plus gros que la normale, et d'apparence vésiculaire en position centrale des fibres musculaires (présenté en figure 3.6). De plus, on observe très fréquemment une augmentation de tissus adipeux et conjonctifs dans les fibres de muscles atteints de CNM (JUNGBLUTH et al., 2018). Concernant les formes de CNM liées à l'X, l'atteinte est présente dès la naissance, tandis que pour les formes non liées à l'X l'atteinte se déclare fréquemment à l'âge adulte, notamment dans le cas de DNM2 (JUNGBLUTH et al., 2018). Au niveau clinique, on retrouve fréquemment les atteintes des muscles extra-oculaires présentes aussi dans les COM, les atteintes bulbaires, respiratoires, cardiaques et des contractures.

3.2.3.4 Myopathies à disproportion congénitale des fibres (CFTD)

Les myopathies à disproportion congénitale des fibres (CFTD) sont un sous-type moins bien défini et spécifique que les trois sous-types présentés précédemment (NM, CNM, COM) qui ont une prévalence d'environ 0,23 pour 100 000. Ce sous-type est principalement défini par la seule présence d'une prédominance des fibres de type 1 et de leur atrophie d'environ 40% par rapport aux fibres de type 2 (présenté en figure 3.7, CLAEYS, 2020). Aucune autre anomalie de structure du muscle n'est présente (tel que les bâtonnets, les cores ou les noyaux centralisés) dans ce sous-type de MC (CLAEYS, 2020). Au niveau clinique, les enfants atteints présentent une hypotonie et une faiblesse musculaire généralisée dès la naissance ou pendant les premières années. De plus, ils présentent une atteinte importante au niveau des muscles du visage et des épaules ainsi que des problèmes respiratoires (CLAEYS, 2020). Ce sous-type de myopathie est lié à une dizaine de gènes tels que ACTA1, MYH7 et RYR1.

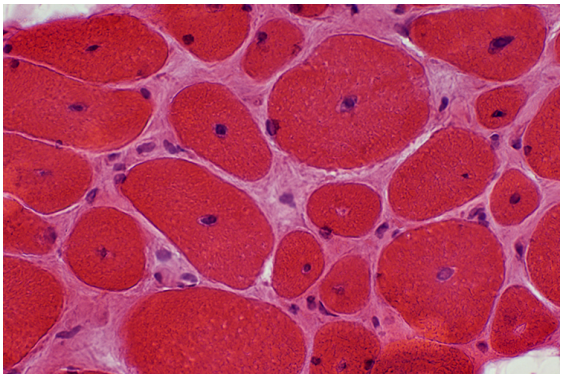


FIGURE 3.6 – **Biopsie de muscle présentant des noyaux centralisés**. Ces noyaux centralisés sont caractéristiques des myopathies centronucléaires à la coloration hématoxyline-éosine (ALAN PESTRONK, 2022).

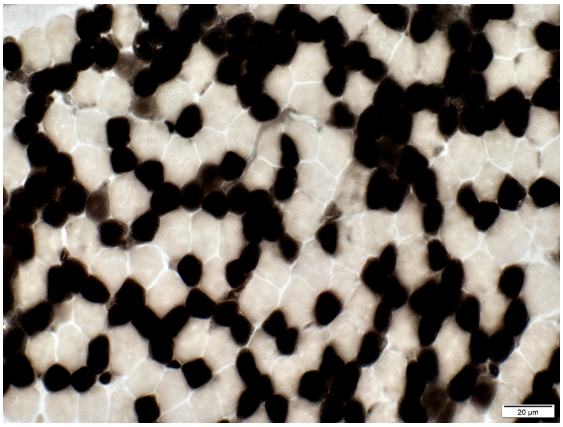


FIGURE 3.7 – **Biopsie de muscle présentant une atrophie et une prédominance des fibres de type 1**. Cette prédominance est caractéristique des myopathies à disproportion congénitale des fibres à la coloration ATPase pH 4.3 (ALAN PESTRONK, 2022).

3.2.3.5 Myopathies de stockage de myosine (MSM)

Enfin, les myopathies de stockage de myosine (MSM), anciennement nommées "*hyaline body myopathy*", se caractérisent principalement par la présence de *"hyaline body"*, des régions d'apparence granulaire et basophile à la coloration hématoxyline-éosine (figure 3.8, Claeys, 2020; Victor Dubowitz et al., 2020). Sur le plan clinique, les patients présentent des problèmes cardiaques, une perte de force distale (mains, poignets), des pieds tombants et une pseudo-hypertrophie des mollets (Cassandrini et al., 2017). Un seul gène est identifié pour l'instant comme pouvant causer une MSM, il s'agit de MYH7, pouvant aussi causer des CFTD et des COM.

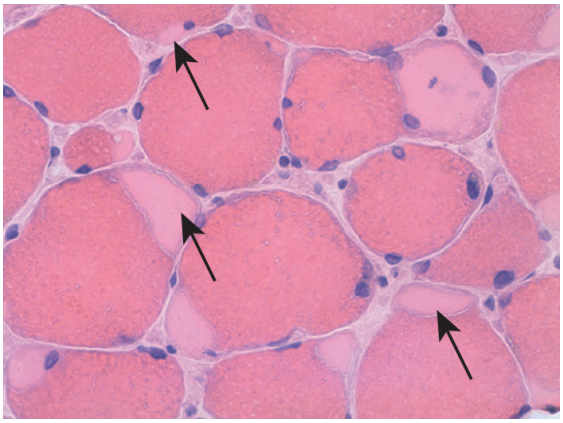


Figure 3.8 – **Biopsie de muscle présentant des *hyaline bodies***. Ces *hyaline bodies* indiqués par les flèches noires sont caractéristiques des myopathies à stockage de myosine (Victor Dubowitz et al., 2020).

La diversité des sous-types de myopathies congénitales tant par leur nombre que par leur hétérogénéité intra-sous-type et les recouvrements sur le plan clinique, histologique et génétique entre les sous-types rendent leur diagnostic difficile. Dans la prochaine section, nous verrons quelles sont les stratégies de diagnostic utilisées actuellement et quelles sont les données générées dans le cadre de ce diagnostic.

3.3 Le défi du diagnostic des myopathies congénitales

Les myopathies congénitales présentent une triple diversité importante sur les plans clinique, histopathologique et génétique. Un sous-type de myopathies peut être causé par une mutation dans différents gènes et une mutation dans un gène spécifique peut causer plusieurs sous-types de myopathies, rendant le diagnostic des patients complexe. De plus, une même mutation peut mener à des manifestations différentes d'un point de vue pathologique (North et al., 2014).

Deux approches pour la classification et le diagnostic des MC existent (North et al., 2014). La première nommée *genotype-up*, consiste à partir de chaque gène responsable de MC et à lister l'ensemble des phénotypes (cliniques ou histopathologiques) suggestifs d'une mutation dans ces gènes. La seconde approchée nommée *"phenotype-down"* consiste, à partir des observations cliniques et histopathologiques, à lister les sous-types de myopathies associés et les potentiels gènes candidats pouvant causer ces phénotypes.

Le processus de diagnostic des MC se base sur trois niveaux d'informations (cliniques, histopathologiques et génétiques), générant ainsi trois types de données (imagerie, textuelles, et

séquences) que nous allons présenter ici pour clore ce chapitre.

3.3.1 Séquençage NGS et données de séquences génétiques

Historiquement, les techniques de séquençage Sanger ont été utilisées pendant des décennies pour identifier les mutations responsables de MC. Cependant, les avancées en termes de séquençage grâce aux séquençage de nouvelle généraiton (*Next Generation Sequencing* (NGS)), permettent aujourd'hui de séquencer beaucoup plus rapidement et à bas cout, l'ensemble d'un panel de gènes d'intérêt pour trouver les mutations présentes dans le génome du patient. Cependant, cette approche permettant d'évaluer un grand nombre de gènes simultanément ne résout pas tous les défis de l'identification des mutations causant les myopathies congénitales. En effet, même avec les techniques de NGS, 50% des patients atteints de myopathies congénitales n'ont pas de diagnostic génétique à ce jour montrant qu'il reste un travail à faire pour l'identification de gènes et de mutations responsables de MC. Par exemple, le projet MYOCAPTURE, porté par le groupe de recherche de Jocelyn Laporte, est issu d'un consortium de sept équipes de recherche avec pour objectif le séquençage de 1000 exomes de familles atteintes de NMD afin d'identifier de nouveaux gènes et mutations.

De plus, comme évoqué précédemment et présenté dans le tableau 3.3 (Cassandrini et al., 2017), un gène peut être responsable de différents sous-types de MC. Les données génétiques seules ne sont donc pas suffisantes pour poser un diagnostic fiable. L'exploitation de ces données de type séquences génétiques requièrent le développement d'outils à même de traiter les résultats de séquençage pour en extraire les mutations, de quantifier la pathogénicité des mutations et de filtrer les mutations pertinentes dans le cadre du diagnostic des MC.

	Genetic determinants										
Clinical and histological characteristics	*ACTA1*	*NEB*	*MTM1*	*MYH7*	*DNM2*	*RYR1*	*TPM2*	*TPM3*	*SEPN1*	*BIN1*	*TTN*
Nemaline myopathy	■	■				■	■	■			
Core myopathy				■		■			■		
Centronuclear myopathy			■		■	■				■	■
Myotubular myopathy			■								
Congenital fiber-type disproportion myopathy (CFTD)	■			■		■	■	■	■		
Myosin storage myopathy				■							

Tableau 3.3 – **Tableau des principaux gènes responsables de myopathies congénitales et des sous-types associés.** Un gène muté peut causer plusieurs types de MC et plusieurs types de MC peuvent être causées par un même gène. (Cassandrini et al., 2017)

3.3.2 Histopathologie et données d'imagerie

L'histopathologie est le premier critère de classification des sous-types de MC. Comme décrit précédemment, plusieurs marqueurs pathologiques sont à évaluer pour pouvoir orienter le diagnostic des MC. Pour observer ces marqueurs, l'examen classique est une biopsie musculaire (pouvant porter sur divers muscles du patient comme le quadriceps, le deltoïde, la vaste externe ou autre) accompagnée d'une observation de la biopsie au microscope sous différentes colorations, fournissant chacune une information sur la structure du muscle. En routine, cinq colorations

principales sont réalisées dans le cadre du diagnostic à partir d'une biopsie musculaire : (i) la coloration Hématoxyline-Eosine (HE) qui révèle la taille des fibres et la position des noyaux (ii) la coloration Trichrome de Gomori (TG) qui révèle la charpente protéique des fibres et qui permet d'observer des agrégats protéiques (iii) la coloration ATPase qui permet de différencier les fibres de type 1 et type 2, et (iv) et (v) les colorations Succinate Déshydrogénase (SDH) et NADH qui révèlent l'activité oxydative des fibres et la position des mitochondries. De plus, des colorations supplémentaires peuvent être réalisées telles que de l'immunohistochimie et les colorations phosphorylases, PAS (glycogène), et Soudan (lipides). Pour finir, dans les cas complexes, il peut être nécessaire d'avoir recours à la microscopie électronique pour observer avec une forte résolution la structure précise des fibres musculaires.

Ainsi, l'examen de la biopsie musculaire génère plusieurs images (une par coloration) de grandes tailles (de l'ordre du millier de fibres par biopsie qui sont évaluées manuellement pour identifier les marqueurs pathologiques. Cependant, comme pour les données génétiques, il est difficile de poser un diagnostic sur la seule base des observations histopathologiques. Comme présenté dans le tableau 3.4 (Jungbluth et al., 2018), les caractéristiques typiques d'un sous-type de myopathies, comme les cores dans les myopathies à cores, peuvent être présents dans d'autres sous-types telles que les myopathies à némaline par exemple. Des travaux de recherche sont encore nécessaires pour trouver des critères réellement discriminants entre les sous-types de myopathies sur le plan histopathologique, tel que le projet de l'Atlas du Muscle porté par Bruno Cadot et Norma Romero (Cadot et Romero, 2022). Cet atlas est une banque de données en ligne d'images d'histopathologie de biopsie de muscle dans différentes colorations pour un large panel de NMD. À ce jour, plus de 5000 images de biopsies musculaires sont disponibles avec le gène et le diagnostic de NMD associés.

Observation	**RYR1 AD**	**RYR1 AR**	**SEPN1**	**TTN**	**MTM1**	**DNM2**	**NEB**	**ACTA1**	**KLHL 40**
Cores	+++	+++	+++	++	-	+	+	+	-
Noyau central	++	++	-	+++	+++	+++	-	-	-
Bâtonnets	+	+	-	+	-	-	+++	+++	+++
CFTD	+	+++	+	+	+	-	-	+	-
Infiltration tissus conjonctifs	++	++	++	+++	-	+	-	-	-

Tableau 3.4 – **Tableau des fréquences des principaux marqueurs pathologiques observables sur la biopsie musculaire en fonction du gène impliqué.** Un gène muté peut provoquer des phénotypes histologiques spécifiques de différents sous-types de myopathies congénitales. Et chaque phénotype histologique peut être causé par plusieurs gènes différents. (Jungbluth et al., 2018).

L'évaluation des données d'imagerie histopathologique est un processus long en raison du nombre de fibres musculaires et du nombre de colorations. En général, il est réalisé de façon qualitative, sans comptage de fibres et des éléments pathologiques.

3.3.3 Comptes rendus cliniques et histopathologiques (données textuelles)

Les observations cliniques de patients sont aussi un niveau d'information utile et nécessaire au diagnostic des myopathies congénitales. Même si certains phénotypes sont très communs et

peu informatifs (hypotonie, faiblesse musculaire générale), d'autres, plus spécifiques, peuvent permettre d'éliminer certains gènes. Le tableau 3.5 (Jungbluth et al., 2018) présente la fréquence des principales observations cliniques en fonction du gène impliqué. On observe par exemple que la présence d'une atteinte cardiaque est très fréquente lors d'une mutation du gène TTN, peu fréquentes dans les gènes ACT1, SEPN1 et RYR1 récessif, et totalement absente pour RYR1 dominant, MTM1, DNM2, NEB et KLHL40. Ainsi il est possible d'orienter les tests génétiques et diagnostiques grâce à cette information. Cependant, on observe aussi que ces phénotypes peuvent apparaitre pour de multiples gènes, rendant impossible le diagnostic sur la seule base de critères cliniques. L'observation des phénotypes cliniques, tout comme l'observation des marqueurs histopathologiques, mène à la rédaction d'un compte rendu clinique textuel qui liste les observations réalisées.

Observation	**RYR1 AD**	**RYR1 AR**	**SEPN1**	**TTN**	**MTM1**	**DNM2**	**NEB**	**ACTA1**	**KLHL 40**
Muscle extra-oculaire	+	+++	-	-	+++	+++	-	-	++
Atteinte bulbaire	+	+++	++	++	+++	++	++	++	+++
Atteinte distale	-	+	-	++	+	+++	++	+	+
Atteinte respiratoire	+	++	+++	++	+++	+	++	++	+++
Atteinte cardiaque	-	+	+	+++	-	-	-	+	-
Contractures	+	+	+	+++	+++	++	++	++	+++

Tableau 3.5 – **Tableau des fréquences des principales observations cliniques en fonction du gène impliqué.** Comme pour les phénotypes histologiques, un même phénotype clinique peut être provoqué par différents gènes mutés et plusieurs phénotypes cliniques différents peuvent être provoqués par un même gène. (Jungbluth et al., 2018).

3.4 Conclusions et synthèse de problématiques

Cette description des myopathies congénitales dresse un portrait complexe du processus de diagnostic, qui requiert l'intégration de trois niveaux d'informations et de modalités différentes (textes, images, séquences). L'ensemble de ces données et leur complexité mettent en avant le besoin d'outils adaptés pour aider à leur exploration automatique. Le but de cette thèse a été de développer des outils capables d'exploiter les comptes rendus et les images de biopsies de patients atteints de myopathies congénitales afin de mieux caractériser les différents sous-types de MC.

Pour les comptes rendus de biopsie, ces documents sont très riches en information sur les patients, cependant ils sont dans un format de texte libre rendant difficile leur utilisation informatique. Afin de pouvoir traiter ces données pour en extraire de nouveaux critères de classification, il est nécessaire de développer des outils adaptés à l'annotation, la structuration et la compréhension de texte libre en langage naturel, notamment grâce aux avancées en IA.

Pour les données d'imagerie, comme mentionné précédemment, en raison de leur volume et taille, leur évaluation est souvent réalisée de façon qualitative uniquement. C'est pourquoi il est intéressant de développer des outils capables de réaliser cette évaluation automatiquement de façon quantitative, grâce à des méthodes IA. Une approche par IA capable de quantifier des éléments sur l'image permet d'extraire des informations plus précises et donc de mieux caractériser et classifier les types de MC, par exemple avec la possibilité d'établir des seuils pour les marqueurs pathologiques.

Développer des méthodes permettant d'exploiter les données d'imagerie (biopsie musculaire) et les données textuelles (cliniques et histopathologies) pour mieux caractériser les MC et aider à leur classification représente deux défis. Le premier, concerne l'explicabilité et la confiance dans les modèles IA, en particulier car il s'agit de données de santé. Il est nécessaire de construire des modèles IA explicables et transparents capables d'extraire de la connaissance à partir des données. Le deuxième défi concerne l'exploitation des données en tant que telles en raison de leur complexité (absence de structure, grande dimensionnalité), il est nécessaire d'utiliser des modèles IA novateurs tels que les réseaux de neurones profonds pour exploiter ces données.

Pour cela, nous avons développé une stratégie d'exploitation des données mélangeant les méthodes ML classiques et les nouvelles technologies IA. Les méthodes ML classiques permettent de construire des modèles transparents et de confiance capable de réaliser des prédictions et d'extraire des connaissances à partir de données bien structurées et annotées. Les nouvelles technologies IA permettent l'extraction d'information de données brutes non structurées, telles que les textes libres et les images. Ainsi en couplant les deux approches il est possible d'utiliser les nouvelles technologies IA comme extracteurs d'information à partir des données brutes de patients (une sorte d'annotation automatique) puis d'appliquer les méthodes ML classiques explicables et transparentes pour réaliser des prédictions automatiques de diagnostic et extraire des informations liant les observations et le diagnostic.

Dans ce cadre-là, à travers le développement d'**IMPatienT** dans le chapitre 5, nous avons créé une application web et base de données permettant l'annotation manuelle des données textuelles et d'imagerie par ontologie pour pouvoir les exploiter par techniques de ML classiques. Dans le chapitre 6, nous avons utilisé différentes techniques de ML explicables pour entrainer des modèles de prédiction à partir de cette base de données. De plus, nous avons développé une méthode d'extraction de connaissances à partir d'un de ces modèles.

Ensuite, dans le chapitre 7, nous avons développé **NLMyo**, un outil basé sur les nouvelles technologies IA pour l'exploitation de texte nommée LLMs pour accélérer et automatiser le processus d'annotation, de référencement et de classification des comptes rendus textuels de patients.

Enfin, dans le chapitre 8, nous avons développé **MyoQuant**, un outil contenant plusieurs méthodes de quantification de marqueurs pathologiques sur les biopsies musculaires de MC, basé sur les nouvelles technologies IA pour l'imagerie. Cet outil a pour objectif, à terme, de permettre la génération d'un compte rendu de biopsie automatique avec des mesures quantitatives des marqueurs.

Deuxième partie

MATÉRIELS ET MÉTHODES

Chapitre 4

Outils informatiques et données utilisées

Dans cette thèse, nous avons développé des méthodes basées sur les IA pour exploiter les données multimodales de patients. Pour cela, nous avons utilisé un vaste panel de ressources biologiques, d'outils informatiques et de méthodes IA. Dans ce chapitre, nous allons décrire l'ensemble des outils et ressources utilisés pour construire nos méthodes d'analyse.

4.1 Données biomédicales de myopathies congénitales

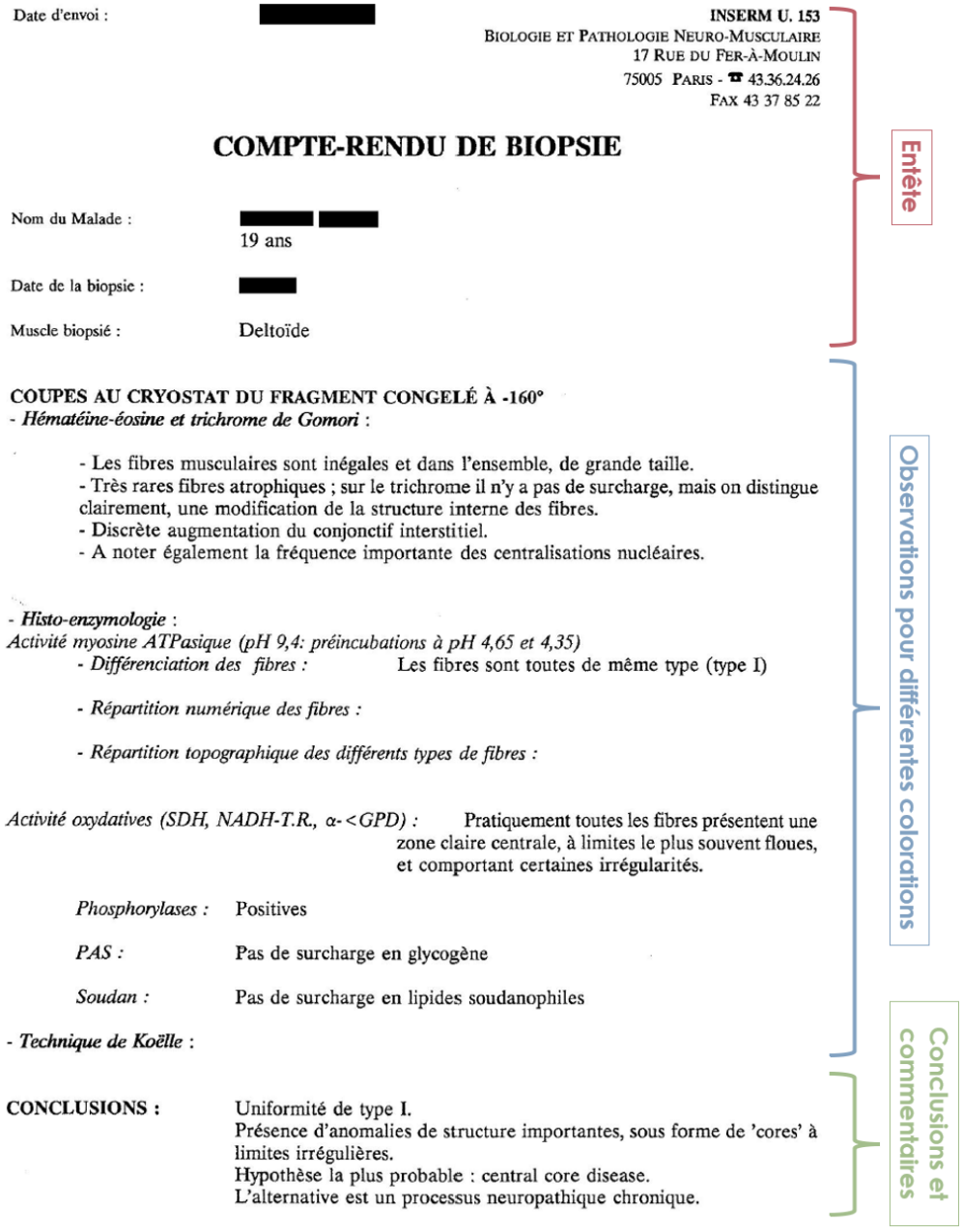
Pour développer ces méthodes, nous nous sommes basés sur des données d'imagerie et des comptes rendus de biopsie. La source de ces données est présentée ci-dessous.

4.1.1 Comptes rendus de biopsie de l'institut de myologie de Paris

La première source de données provient de l'institut de myologie de Paris. Grâce à une collaboration avec l'équipe du laboratoire d'histopathologie d'abord dirigé par Norma B. Romero puis Teresinha Evangelista, nous avons pu récupérer et utiliser 192 comptes rendus de biopsie musculaire de patients atteints de myopathies (congénitales, dystrophies ou autre), dont 138 spécifiquement atteint par des myopathies congénitales identifiées. Ces rapports sous format papier ont été scannés puis anonymisés d'abord avec un outil d'anonymisation que nous avons développé (présenté dans le chapitre 7), puis vérifiés à la main. La figure 4.1 présente la structure d'un compte rendu anonymisé de biopsie typique présent dans le jeu de données. Il y a deux types de comptes rendus, ceux qui concernent les observations en microscopie photonique et ceux qui concernent les observations en microscopie électronique. Cependant, cette structure peut varier en fonction de l'année de production du compte rendu, certains sont totalement déstructurés.

4.1.2 Images de biopsie musculaire de souris

Une seconde source de données provient d'une collaboration avec l'Institut de génétique et de biologie moléculaire et cellulaire (IGBMC), plus spécifiquement avec l'équipe Physiopathologie des maladies neuromusculaires dirigée par Jocelyn Laporte. Cette équipe travaille sur les



Date d'envoi : [redacted]

INSERM U. 153
BIOLOGIE ET PATHOLOGIE NEURO-MUSCULAIRE
17 RUE DU FER-À-MOULIN
75005 PARIS - ☎ 43.36.24.26
FAX 43 37 85 22

COMPTE-RENDU DE BIOPSIE

Nom du Malade : [redacted] [redacted]
19 ans

Date de la biopsie : [redacted]

Muscle biopsié : Deltoïde

COUPES AU CRYOSTAT DU FRAGMENT CONGELÉ À -160°
- ***Hématéine-éosine et trichrome de Gomori*** :

- Les fibres musculaires sont inégales et dans l'ensemble, de grande taille.
- Très rares fibres atrophiques ; sur le trichrome il n'y a pas de surcharge, mais on distingue clairement, une modification de la structure interne des fibres.
- Discrète augmentation du conjonctif interstitiel.
- A noter également la fréquence importante des centralisations nucléaires.

- ***Histo-enzymologie*** :
Activité myosine ATPasique (pH 9,4: préincubations à pH 4,65 et 4,35)
- *Différenciation des fibres :* Les fibres sont toutes de même type (type I)

- *Répartition numérique des fibres :*

- *Répartition topographique des différents types de fibres :*

Activité oxydatives (SDH, NADH-T.R., α-<GPD) : Pratiquement toutes les fibres présentent une zone claire centrale, à limites le plus souvent floues, et comportant certaines irrégularités.

Phosphorylases : Positives

PAS : Pas de surcharge en glycogène

Soudan : Pas de surcharge en lipides soudanophiles

- ***Technique de Koëlle*** :

CONCLUSIONS : Uniformité de type I.
Présence d'anomalies de structure importantes, sous forme de 'cores' à limites irrégulières.
Hypothèse la plus probable : central core disease.
L'alternative est un processus neuropathique chronique.

FIGURE 4.1 – **Exemple de compte rendu de biopsie en microscopie photonique anonymisé de l'institut de myologie de Paris**. Les comptes rendus sont structurés en trois sections : un entête avec des informations générales sur la patient, un corps de texte contenant une liste de colorations et les observations réalisées et une conclusion avec le diagnostic final et un commentaire optionnel.

myopathies congénitales et utilise plusieurs souris modèles de myopathies congénitales. Ainsi en travaillant avec les membres de l'équipe réalisant des biopsies musculaires sur ces modèles, nous avons développé des méthodes d'analyse pour des biopsies musculaires de souris aux colorations HE, SDH, ATPase et à fluorescence.

4.2 Ontologies et nomeclatures en biologies

En biologie, les ontologies sont des vocabulaires standards pour faciliter l'intégration des données et leur analyse. Dans cette thèse, pour standardiser les données issues des comptes rendus, notamment dans le cadre du développement de l'outil **IMPatienT** et l'analyse de sa base de données, nous avons utilisé diverses ontologies pré-existantes que nous allons décrire ici.

4.2.1 Ontologie des phénotypes : HPO

L'ontologie *Human Phenotype Ontology*, HPO, développée en 2008 par Peter N Robinson et Sebastian Köhler au *Charité University Hospital* à Berlin (Robinson et al., 2008, Köhler et al., 2021), rassemble l'ensemble des phénotypes médicaux observables chez l'Homme. Organisée sous forme d'arbre, elle contient plus de 13 000 termes organisés selon un niveau croissant de précision (par exemple le terme "anomalie de l'œil" est un parent du terme "anomalie de la pupille"). Chaque terme est associé à un identifiant unique sous la forme HPO :XXXXX et possède un certain nombre d'annotations comme des maladies associées, des gènes associés, des synonymes, des publications associées. L'ensemble de ces informations est disponible en ligne sur le portail https://hpo.jax.org/ qui permet aussi de télécharger l'ontologie dans les formats standards (JSON, OBO, OWL). Cette ontologie est utilisée dans **IMPatienT** pour normaliser les observations cliniques des patients.

4.2.2 Ontologie de maladies : ORDO par Orphanet

L'*Orphanet Rare Disease Ontology*, ORDO est développée dans le cadre d'une collaboration entre Orphanet (https://www.orpha.net/, Maiella et al., 2013), et l'Institut Européen de Bioinformatique (EBI). Orphanet est une ressource informatique ayant pour but de répertorier l'ensemble des informations concernant les maladies rares et les médicaments orphelins. L'ontologie ORDO répertorie plus de 7 000 maladies rares connues. Chaque maladie rare est répertoriée sous un identifiant unique de la forme ORPHA :XXXXXX, par exemple la maladie "Myopathie congénitale sévère à némaline" correspond à l'identifiant ORPHA :171430. De plus, chaque maladie est associée à des annotations telles que leur prévalence, des synonymes, un mode d'hérédité, un âge d'apparition, un pronostic, les gènes causant la maladie et autre. Pour finir, chaque maladie est aussi liée à des symptômes cliniques grâce à un lien direct vers des identifiants de l'ontologie HPO. Dans le cadre de notre outil **IMPatienT** nous avons utilisé cette ontologie pour normaliser le diagnostic final des patients.

4.2.3 Nomenclature génétique : HGNC et HGVS

La nomenclature HUGO (https://www.genenames.org/, acronyme de Human Genome Organisation, est gérée par le Comité de Nomenclature des Gènes de HUGO (HGNC) à l'Institut Européen de Bioinformatique. Ce comité est responsable de l'attribution de noms uniques pour les gènes humains, que ce soit des gènes codants pour des protéines, gènes non codants ou pseudogènes. Au total, plus de 43 000 noms de gènes uniques sont référencés et annotés avec des références croisées vers des bases de données externes (banque de séquences, orthologies,

mutations, structures, Orphanet...). Concernant les variations génétiques (mutations), la nomenclature établie par la *Human Genome Variation Society*, HGVS (https://www.hgvs.org/) fait autorité. Cette nomenclature spécifie la façon de représenter textuellement un variant génétique. Par exemple selon cette nomenclature la notation "NM_001164508.2(NEB) :c.25336C>T (p.Arg8446Ter)" indique une mutation faux-sens dans la protéine NEB où l'acide aminé n° 8446 est substitué d'une arginine à un codon stop. La nomenclature HUGO et HGVS sont toutes deux intégrées dans **IMPatienT** (voir chapitre 5) pour codifier le diagnostic génétique des patients (gène muté et localisation de la mutation responsable de la maladie).

4.3 Développement de modèles de ML traditionnels et xAI

Dans le cadre de l'analyse de la base de données de **IMPatienT** et des résultats d'*embedding* de **NLMyo** nous avons utilisé des algorithmes de ML traditionnels pour créer des modèles prédictifs. Dans cette section, nous présentons les principaux outils utilisés en ce qui concerne les algorithmes utilisés et l'analyse de leurs performances.

4.3.1 *Scikit-Learn* : une boite à outils pour l'apprentissage automatique

Scikit-Learn (version 1.3, Pedregosa et al., 2011, https://scikit-learn.org/) est une bibliothèque de code python mettant à disposition un grand nombre d'outils et permettant de préparer les données, d'entrainer des modèles de *ML* et d'évaluer ses performances. *Scikit-learn* inclut des algorithmes de classification, de régression, de clustering, de réduction de dimensionnalité, d'optimisation et de sélection de modèles. Scikit-learn a été utilisé dans cette thèse pour : (i) partitionner et normaliser les données, (ii) entrainer des modèles de classification de différentes familles d'algorithmes, (iii) optimiser et évaluer les performances des modèles.

4.3.2 Validation croisée et évaluation des performances

La validation croisée (*cross-validation* en anglais) est une technique d'évaluation de modèles prédictifs permettant d'avoir une estimation plus robuste et précise des métriques de performance. Cette méthode est particulièrement utile pour les jeux de données de petite taille. La figure 4.2 présente schématiquement son fonctionnement. Le jeu de données initial est séparé en N sous-ensembles (nommés *folds* ou bloc, ici au nombre de 5). Ensuite le modèle est entrainé sur l'ensemble des sous-ensembles à l'exception d'un, qui est utilisé comme jeu de test des performances. Ce processus est répété autant de fois qu'il y a de *folds* afin que chaque *fold* ait été utilisé exactement une fois comme jeu de test. Ainsi pour 5 *folds*, cinq modèles sont entrainés et évalués. On peut alors calculer une performance moyenne du modèle entrainé sur l'ensemble des données, en calculant les performances moyennes des cinq modèles. Plus le nombre de *folds* est important, plus cette moyenne est précise, mais c'est plus coûteux en temps et ressources de calcul, car il faut entrainer plus de modèles.

4.3.3 Recherche d'hyperparamètres

La recherche d'hyperparamètres est une étape importante dans le développement de modèles prédictifs pour améliorer les performances des modèles. Les hyperparamètres sont des paramètres propres à chaque algorithme, qui sont spécifiés en amont de l'entrainement et qui ne varient pas lors de l'entrainement, mais qui influent directement sur les performances du

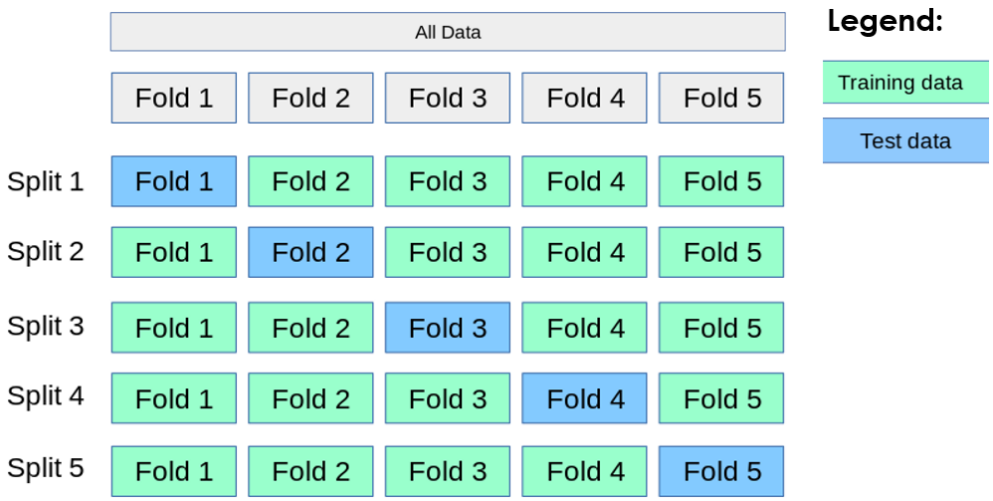


FIGURE 4.2 – **Schéma du principe de la validation croisée**. Ici la valiation-croisée est réalisée en *5 folds*, chaque point de donnée sera utilisé une fois dans le jeu de test et 4 fois dans 4 jeux d'entrainement (modifié de la documentation de *Scikit-Learn*).

modèle. Par exemple, dans le cadre de l'entrainement d'un arbre de décision, un exemple d'hyperparamètre peut être la profondeur maximale de l'arbre ou encore la méthode de calcul de qualité d'un nœud (par exemple méthode de gini ou entropie).

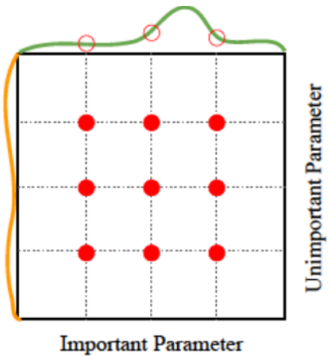


FIGURE 4.3 – **Schéma du principe de recherche d'hyperparamètres par grille pour 2 paramètres**. Un point rouge représente un modèle entrainé. Les courbes en abscisse et ordonnée représente les performances du modèle. Le paramètre en ordonnée n'a que très peu d'effet sur les performances, alors que les variations du paramètre en abscisse améliorent les performances du modèle.

La recherche d'hyperparamètres revient à trouver une combinaison de paramètres optimale qui maximise les performances du modèle après entrainement. La méthode la plus classique pour cela est l'optimisation par grille. À partir d'un ensemble de valeurs possibles pour chaque paramètre, on va entrainer et tester les performances du modèle pour chaque combinaison de valeurs et sélectionner la combinaison de valeurs la plus performante. La figure 4.3 présente un exemple théorique d'optimisation par grille pour deux paramètres. Pour chacun des paramètres, 3 valeurs sont possibles, c'est donc un total de 9 modèles qui sont entrainés et testés. On

peut alors mesurer l'impact de chaque paramètre sur les performances finales du modèle pour trouver le paramètre le plus important et sa valeur optimale. D'autres méthodes moins naïves et moins coûteuses que la grille existent pour la recherche d'hyperparamètres telle que l'approche bayésienne utilisée par la bibliothèque de code *Optuna* (Akiba et al., 2019, https://optuna.org/) qui permet de trouver une combinaison optimale plus rapidement.

4.3.4 Algorithme de système de classeurs : ExSTraCS 2.0

Les systèmes de classeurs sont une famille d'algorithmes de ML considérés comme explicables. Dans cette thèse, nous avons utilisé le LCS nommé *ExSTraCS 2.0* (R. J. Urbanowicz et Moore, 2015), spécifiquement développé pour les tâches de classification à partir de données complexes, hétérogènes et de haute dimensionnalité. Cet algorithme a été utilisé pour tenter de prédire le diagnostic de sous-type de myopathies congénitales des patients à partir des annotations réalisées dans **IMPatienT**. L'implémentation en Python de cet algorithme de LCS nommée *scikit-ExSTraCS* (https://github.com/UrbsLab/scikit-ExSTraCS) nous a permis d'entrainer un modèle basé sur cette méthode et de comparer ses performances aux autres algorithmes implémentés dans *scikit*, *via* le *pipeline* d'entrainement et de comparaison de modèles nommé *Streamline*.

4.3.5 Streamline : un *pipeline* d'entrainement et de comparaison de modèles ML

Pour entrainer le modèle de classification le plus performant possible dans le cadre de l'analyse de la base de données d'**IMPatienT**, nous avons utilisé le *pipeline* d'entrainement et de comparaison de modèles de ML nommé *Streamline* (R. Urbanowicz et al., 2023). La figure 4.4 (R. Urbanowicz et al., 2023) présente son fonctionnement. Il y a 4 étapes principales. L'étape 1 (modules 1 et 2) consiste en la lecture, exploration statistique et préparation des données. L'étape 2 (modules 3 et 4) consiste au calcul de l'importance de chaque annotation (comprendre ici : *feature*) et en la sélection des annotations les plus pertinentes pour la classification. L'étape 3 (module 5) consiste à l'optimisation et l'entrainement de 16 algorithmes de ML variés et l'évaluation des performances des modèles issus de cet entrainement. Finalement, l'étape 4 (modules 6, 7, 8 et 9) correspond à la phase de post-traitement, où seront générés les figures de comparaison des performances des modèles et le nettoyage des fichiers. STREAMLINE n'est pour l'instant disponible que pour faire de la classification binaire. Nous nous intéressons à un problème de classification multi-classe (prédiction de diagnostic), donc nous avons modifié le pipeline pour le rendre compatible avec nos données. Le code du pipeline modifié est accessible à l'adresse https://github.com/lambda-science/STREAMLINE.

4.3.6 Métriques de performance

Pour évaluer les performances de nos modèles, nous avons utilisé plusieurs métriques. Dans le cadre d'une classification multi-classe et déséquilibrée, la mesure de l'exactitude de classification (*accuracy*) n'est pas suffisante pour obtenir une bonne représentation des performances des modèles. Ainsi nous mesurons aussi l'exactitude équilibrée (*balanced accuracy*), le score F1, score F1 macro, la spécificité et le coefficient de corrélation de Matthew. L'ensemble de ces métriques de performance, décrites dans la section suivante, sont calculées à partir de la matrice de confusion.

4.3.6.1 Matrice de confusion

La matrice de confusion est une matrice en deux dimensions qui compare les prédictions d'un modèle par rapport aux labels réels des données. Par exemple pour une classification en 3 classes,

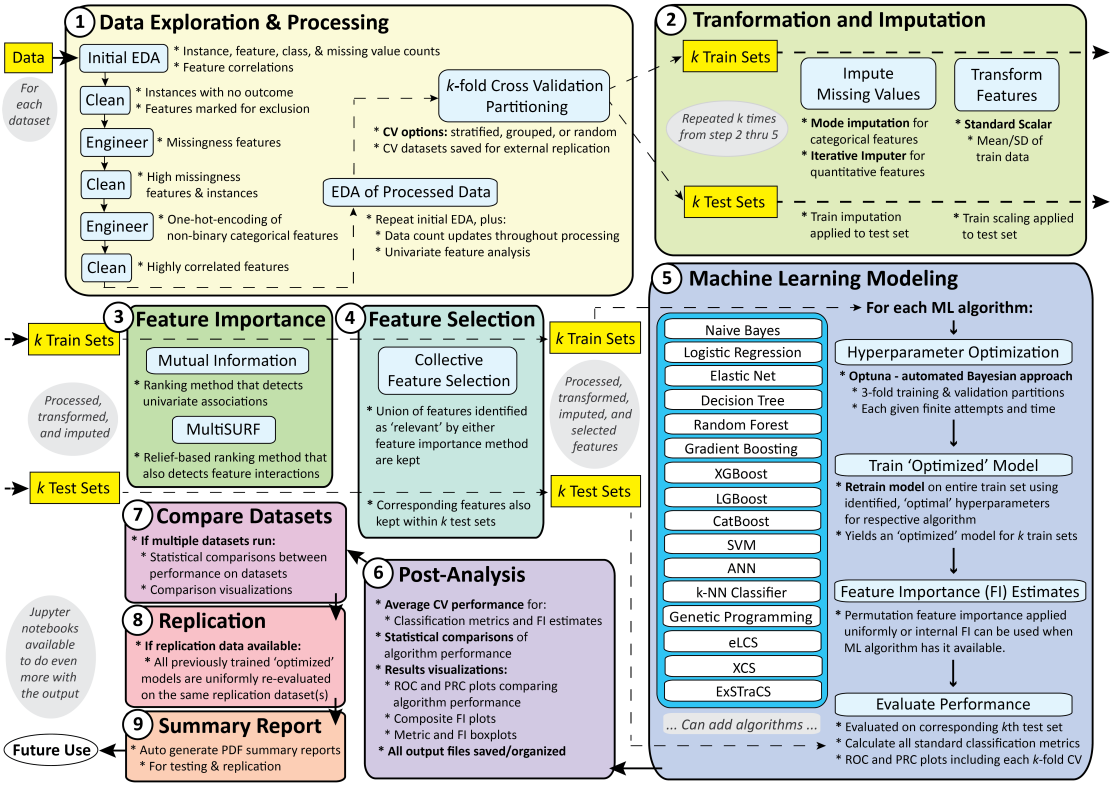


FIGURE 4.4 – **Schéma du pipeline STREAMLINE**. Ce pipeline est divisé en 4 parties principales : (i) traitement des données, (ii) calcul de l'importance et sélection des descripteurs, (iii) entrainnement et optimisation des modèles, (iv) post-traitement et comparaison des modèles. (R. URBANOWICZ et al., 2023)

une matrice de confusion de taille (3x3) est produite. La diagonale (haut gauche vers bas droit) représente l'ensemble des prédictions correctes (vrais positifs et vrais négatifs) c'est-à-dire les points pour lesquels la prédiction est en accord avec le label réel. Les autres cases représentent l'ensemble des prédictions incorrectes (faux positifs ou faux négatifs). La figure 4.5 représente un exemple de matrice de confusion pour une classification binaire en 2 classes.

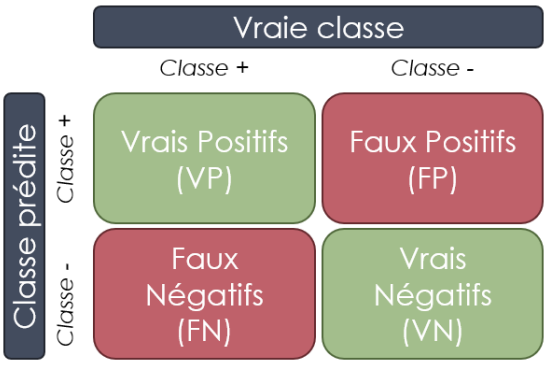


FIGURE 4.5 – **Exemple de matrice de confusion binaire.** Pour une classification à deux classes la matrices de confusion est composée de vraispositif (VP), vrais négatifs (VN), faux positifs (FP), faux négatifs (VN).

4.3.6.2 Exactitude

L'exactitude est une mesure de performance classique qui évalue la proportion de points de données correctement classées par rapport aux nombres totaux de points de données. Cette mesure est trompeuse dans le cadre de données déséquilibrées. Elle est calculée telle que :

$$\text{Exactitude} = \frac{\text{Vrais Positifs} + \text{Vrais Négatifs}}{\text{Vrais Positifs} + \text{Vrais Négatifs} + \text{Faux Positifs} + \text{Faux Négatifs}}$$

4.3.6.3 Exactitude équilibrée

L'exactitude équilibrée quant à elle est utile dans le cadre d'un jeu de données déséquilibré. Elle donne une importance égale aux performances de chaque classe. Pour cela, l'exactitude (équivalente à la sensibilité dans un contexte multi-classe) est calculée pour chaque classe. Ensuite, l'exactitude équilibrée correspond donc à la moyenne de l'exactitude pour chaque classe. Ainsi sa formule est égale à :

$$\text{Exactitude équilibrée} = \frac{1}{n}\sum_{i=1}^{n} \text{Exactitude}_i$$

où n représente le nombre de classes différentes.

4.3.6.4 Précision, sensibilité, spécificité et Score F1

La figure 4.6 présente les notions de précision, sensibilité (rappel) et spécificité en prenant comme exemple l'interprétation d'un test COVID. Pour un test COVID, la précision représente la proportion de patients réellement positifs parmi des tests positifs. Le sensibilité (ou rappel)

représente la proportion de tests positifs parmi l'ensemble des personnes positives à la COVID. Enfin la spécificité mesure la proportion de tests réellement négatifs parmi l'ensemble des patients négatifs au COVID.

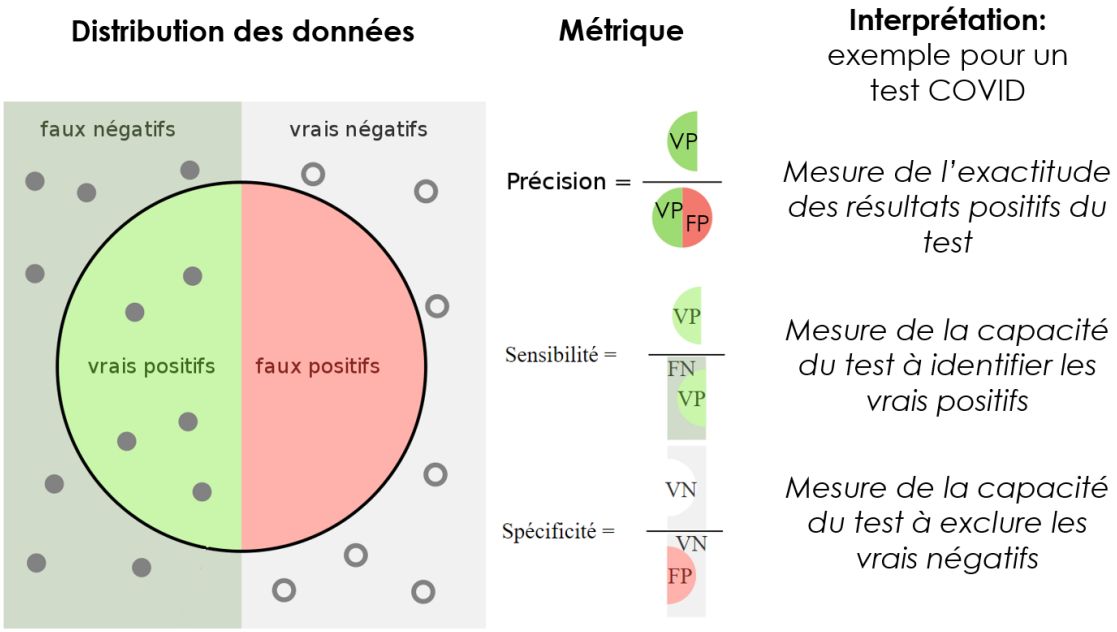


FIGURE 4.6 – **Schéma de la notion de précision, sensibilité et spécificité**. L'exemple d'un test de dépistage COVID est utilisé pour illustrer l'utilité des différentes mesures pour évaluer les performances du test (modifiée de Wikipédia "Précision et rappel").

Le score-F1 correspond à la moyenne harmonique de la précision et de la sensibilité (rappel) et donc permet de tenir compte à la fois des faux positifs et des faux négatifs. Dans le cadre de classification multi-classe (3 et plus), le score F1 peut être calculé de plusieurs manières. Soit de manière "micro" c'est-à-dire globalement à partir du nombre de vrais positifs, faux négatifs et faux positifs. Soit de manière "macro", en calculant le score F1 de chaque classe et en réalisant leur moyenne, similairement à la différence entre exactitude et exactitude équilibrée. Ainsi la formule du score F1-Macro est :

$$\text{Score F1 Macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{2\times \text{Précision}_i \times \text{Rappel}_i}{\text{Précision}_i + \text{Rappel}_i}$$

où n représente le nombre de classes différentes.

4.3.6.5 Coefficient de corrélation de Matthew

Le coefficient de corrélation de Matthew *(Matthew Correlation Coefficient, MCC)* est une métrique prenant en compte l'ensemble des éléments de la matrice de confusion (vrais positifs, faux positifs, vrais négatifs, faux négatifs), contrairement aux métriques présentées ci-dessus. De plus, elle est équilibrée, c'est-à-dire qu'elle n'est pas biaisée dans le cas de classes déséquilibrées. Ses valeurs sont comprises entre -1 et 1, avec 0 pour des prédictions aléatoires, 1 pour des prédictions parfaites et -1 pour des prédictions parfaitement contraires. Cette métrique est plus

informative sur la qualité d'un modèle que le score F1 (Chicco et Jurman, 2020). Dans le cadre d'une classification binaire, la formule du MCC est :

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$$

Dans le cadre d'une classification multi-classe, la formule est plus complexe :

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

où c représente le nombre d'échantillons correctement prédits, s représente le nombre total d'échantillons, p_k le nombre de fois où la classe k a été prédite, t_k le nombre de fois où la classe k s'est réellement produite.

4.4 Techniques d'analyse d'image et réseaux de neurones

Les méthodes que nous avons développées permettent d'analyser des données d'imagerie histologique. L'outil **IMPatienT** présenté dans le chapitre 5 permet de faire de l'annotation et segmentation d'image en utilisant des techniques d'analyse d'image traditionnelles. L'outil **MyoQuant** présenté dans le chapitre 8 permet de faire de la quantification de marqueurs pathologiques grâce à la fois à des méthodes d'analyse traditionnelles, mais aussi grâce à des modèles pré-entrainés et des réseaux de neurones profonds. Dans cette section nous allons voir les bibliothèques de codes, les modèles et le matériel que nous avons utilisés.

4.4.1 Méthodes d'analyse d'image traditionnelles avec scikit-image

Pour l'analyse d'image en utilisant des méthodes traditionnelles nous avons utilisé la bibliothèque de code *scikit-image* (Walt et al., 2014, https://scikit-image.org/). *Scikit-image* a été développé en 2014 et met à disposition des outils de base pour l'analyse d'image tel que le calcul de contraste, d'intensité et de texture de pixels qui sont des mesures utilisées par **IMPatienT** pour la segmentation d'image. Dans **MyoQuant**, *scikit-image* est utilisé pour tracer des lignes, réaliser de l'érosion d'image et mesurer des surfaces, périmètres, diamètre de Feret et la position des centroïdes à partir de masques de segmentation.

4.4.2 Modèle pré-entrainé Cellpose et Stardist

Dans le cadre de **MyoQuant** nous avons utilisés des modèles pré-entrainés pour l'analyse des images de coupes histologiques. Pour la segmentation des fibres musculaires, nous avons utilisé l'implémentation Python du modèle Cellpose (Stringer et al., 2021, https://github.com/MouseLand/cellpose). Parmi les différents modèles inclus dans Cellpose, nous avons utilisé spécifiquement le modèle *cyto2*, qui est le modèle disponible dans Cellpose le plus récent et performant pour la segmentation des fibres musculaires sous diverses colorations.

Quant à la segmentation des noyaux cellulaires, nous avons utilisé l'implémentation Python du modèle Stardist (Weigert et al., 2020, https://github.com/stardist/stardist). Plus précisément dans le cadre de l'analyse des noyaux dans la coloration HE, nous avons utilisé le modèle pré-entrainé *2D_versatile_he*, qui est un modèle spécifiquement entrainé sur des images à la coloration HE.

4.4.3 Développement de réseaux de neurones profond de type ResNet avec Keras Tensorflow

En plus des modèles pré-entrainés, dans **MyoQuant** nous avons entrainé et intégré nos propres modèles de classification basés sur les réseaux de neurones profonds. Pour cela nous avons utilisé les bibliothèques de code Keras (Chollet et al., 2015, https://keras.io/) et Tensorflow (Martín Abadi et al., 2015, https://www.tensorflow.org/). Keras est une bibliothèque qui permet une interaction simplifiée avec Tensorflow mettant à disposition plusieurs architectures de modèles préétablis ainsi que des fonctions permettant d'accélérer le développement de modèles de réseaux de neurones. Dans le cadre de développement de modèles de classification d'image, nous avons utilisé l'architecture RestNet50 version 2 pré-entrainée sur le jeu de données *ImageNet* implémentée dans Keras. L'architecture ResNet50 (He et al., 2015) est un réseau de neurones convolutifs composé de 48 couches convolutives. Ce modèle possède un total de 23 564 800 paramètres. L'entrainement de ce modèle et son inférence sur des données d'imagerie ont été réalisés sur des machines virtuelles de la plateforme SCIGNE Grand-Est équipée de GPU RTX 2080 Ti.

4.5 Développement d'outils basés sur modèles linguistiques de grande taille

Les méthodes que nous avons développées (**NLMyo**) pour analyser du texte libre (compte rendu de biopsie) se basent sur les modèles linguistiques de grande taille. Dans cette section, nous détaillons les outils et modèles que nous avons utilisés.

4.5.1 Reconnaissance de texte avec Tesseract

Dans un premier temps, comme la majorité des comptes rendus sont au format PDF il est nécessaire de les convertir en texte. Pour cela, nous avons utilisé des méthodes d'reconnaissance optique de caractères (*Optical Character Recognition*, OCR). L'outil libre que nous avons utilisé pour cela se nomme Tesseract (Ray et al., 2015) version 5, mise à disposition en novembre 2021. Cet outil est capable de reconnaitre de manière robuste du texte dactylographié dans plus de 100 langues différentes.

4.5.2 Modèles linguistiques de grande taille utilisés

Dans le cadre du développement de **NLMyo** nous avons utilisé 2 LLMs génératifs et 2 LLMs d'embedding. Pour chaque catégorie de LLMs nous avons voulu comparer les résultats issus de modèles provenant de fournisseurs externes (OpenAI) par rapport à des modèles plus petits hébergés localement.

4.5.2.1 Modèles génératifs : OpenAI GPT-3.5 et Vicuna-7B

Un des critères déterminants pour le choix de modèle génératif est les performances du modèle et la taille de contexte. La taille de contexte pour un LLMs représente le nombre de mots qu'il est capable de traiter lors d'une requête. Ainsi plus la taille de contexte est grande, plus il est possible d'analyser un document de grande taille avec des instructions détaillées. Par exemple, le modèle GPT-3.5-turbo d'OpenAI a une taille de contexte de 4096 jetons, ce qui représente environ 3000 mots en anglais.

La grande majorité des modèles open source et auto-hébergeable ont une taille de contexte de 512, ce qui limite leur utilisation pour l'analyse de grands documents. Notre choix de modèle auto-hébergeable s'est porté sur le modèle Vicuna-7B-1.1 (Chiang et al., 2023, https://huggingface.co/vicuna/ggml-vicuna-7b-1.1), qui en plus d'être un modèle de petite taille (7 milliards de paramètres par rapport aux 175 milliards de paramètres de GPT-3.5-turbo) donc requérant de faible ressources informatiques, possède une taille de contexte de 2048. De plus ce modèle est actuellement le plus performant parmi les modèles open source de taille 7B (Hendrycks et al., 2021).

Concernant le modèle provenant d'un fournisseur externe, nous avons choisi d'utiliser GPT-3.5-turbo d'OpenAI, car ce modèle allie à la fois une grande taille de contexte de 4096, d'excellentes performances (4e modèle le plus performant toutes catégories confondues Lianmin Zheng et al., 2023) et possède une API accessible à bas coût (0,002 $ par tranche de 1000 *tokens*).

Pour finir, pour les deux modèles génératifs, nous avons mis le paramètre de température le plus proche de 0 possible, c'est-à-dire à 0,01. Le paramètre de température contrôle le niveau de hasard des réponses du modèle. Plus ce paramètre est proche de 0, plus la réponse du modèle est déterministe. Plus cette valeur est supérieure à 1, plus la sortie est aléatoire. La valeur par défaut du modèle est 1. Ainsi en mettant ce paramètre très proche de 0, les résultats sont reproductibles.

4.5.2.2 Modèle d'*embedding* : OpenAI et Instructor

Les LLMs d'embedding sont des modèles qui transforment du texte en vecteur numérique, ce qui permet de faire de la classification, du clustering et de la recherche de similarité. Ces modèles sont utilisés dans **NLMyo** pour faire de la prédiction de diagnostic et pour créer un moteur de recherche de patients.

Comme pour les modèles génératifs, nous avons utilisé deux modèles, un par un fournisseur externe et un auto-hébergé pour comparer les performances. En termes de choix de modèles, pour le modèle issu de fournisseurs externes, nous avons utilisé le modèle d'*embedding* nommé text-embedding-ada-002, car nous utilisons déjà le modèle génératif du même fournisseur (OpenAI). De plus ce modèle est multilangue et performant, car il est classé 6e en termes de performances d'*embedding* sur 75 modèles testés à travers un panel de 56 jeux de données (Muennighoff et al., 2022).

Concernant le modèle auto-hébergé, nous avons choisi d'utiliser le modèle nommé Instructor (Su et al., 2023, https://huggingface.co/hkunlp/instructor-large). Ce modèle est parmi les plus performants, classé 2e sur 75 modèles testés (Muennighoff et al., 2022), de plus il comporte dans son jeu d'entrainement des données issues de textes médicaux, ce qui permet d'espérer de bonnes performances pour nos comptes rendus de biopsies.

4.5.3 Interaction avec les modèles linguistiques de grande taille avec LangChain

Face à la diversité des LLMs génératifs, LLMs d'*embedding* et des outils associés, il est nécessaire d'avoir un outil qui uniformise la façon d'interagir avec ces modèles. C'est l'objectif de la bibliothèque de code nommée LangChain (Chase Harrison, 2022) que nous avons utilisé à travers **NLMyo**. Cette bibliothèque de code permet d'unifier la façon d'interagir avec les modèles auto-hébergés et les différents fournisseurs externes, ce qui permet d'accélérer la phase d'exploration des performances des modèles et le développement d'outils basés sur les LLMs.

En plus des interactions avec les modèles de langage, LangChain met à disposition des outils pour interagir avec des bases de données optimisées pour le stockage et la requête de vecteurs numériques (résultats des modèles d'*embedding*). Ainsi dans **NLMyo** nous avons utilisé la base de

données de vecteurs nommée ChromaDB (https://www.trychroma.com/). Les bases de données de vecteurs sont des bases de données qui permettent de stocker des documents ainsi que leurs résultats d'*embedding*. Ces bases sont optimisées pour le calcul de similarité entre un vecteur requête et une grande base de données de vecteurs. Cela permet par exemple de construire des moteurs de recherche de documents, dans notre cas il s'agit d'un moteur de recherche de comptes rendus de biopsie requêtables par symptômes.

4.6 Développement d'outils et d'interfaces

Les divers outils que nous avons développés dans cette thèse (**IMPatienT**, **NLMyo**, **MyoQuant**) sont disponibles sous différentes formes : applications web complètes, démonstrations en ligne, outils en ligne de commande. Pour cela, nous avons utilisé diverses bibliothèques de code spécifique à chaque cas de figure.

4.6.1 Développement d'une application web complète pour IMPatienT

Dans le cadre du développement de notre application web et base de données **IMPatienT** nous avons entièrement construit une application web. Les applications web sont composées de trois éléments : l'interface (nommée *front-end*), le serveur de calcul (nommé *back-end*) et la base de données. Pour construire l'interface de **IMPatienT**, nous avons utilisé la bibliothèque de code graphique CSS BootStrap 5 (Mark Otto et Jacob Thornton, 2011) couplée à la bibliothèque de code JavaScript JQuery 3.7 (Resig, 2006) pour l'interactivité du site. Concernant la partie serveur, nous avons utilisé la bibliothèque de code Python nommée Flask (Ronacher, 2010). Nous avons utilisé Flask car c'est une bibliothèque minimale et bien documentée qui permet de construire un site web rapidement en augmentant graduellement la complexité. Enfin, en guise de base de données, nous avons utilisé le système de base de données relationnelle SQLite Hipp, 2020. Le système SQLite a l'avantage d'être léger et portable, car la base de données n'est composée que d'un seul fichier tout en proposant presque les mêmes fonctionnalités que les systèmes de base de données relationnelles plus complexes.

4.6.2 Développement d'un outil en ligne de commande pour MyoQuant

Dans le cadre du développement de notre outil d'analyse d'images **MyoQuant**, nous avons décidé de rendre l'outil disponible d'abord sous la forme d'outil en ligne de commande. En effet, comme **MyoQuant** est voué à analyser des images de grande taille pendant des temps d'analyse longs et sur des ordinateurs avec de l'équipement spécialisé, l'outil en ligne de commande semble être la modalité la plus adaptée. Pour cela nous avons utilisé les bibliothèques de code python nommées Typer (Ramírez, 2019) et Rich (Will McGugan, 2020). Typer permet de créer des commandes complexes pour un outil en ligne de commande, de valider les paramètres et de générer automatiquement la documentation nécessaire à l'outil, tandis que Rich permet d'enrichir l'expérience de l'usage de la ligne de commande en affichant des données formatées directement dans le terminal (en couleur, sous forme de tableau, interactives etc.).

4.6.3 Développement de démonstrations en ligne pour NLMyo et MyoQuant

Enfin, pour le développement de **NLMyo** mais aussi pour **MyoQuant** nous avons voulu développer des applications de démonstration en ligne. Ces applications de démonstration ont pour but de faciliter la communication autour des outils et de pouvoir démontrer leur utilité. Pour cela nous avons utilisé la bibliothèque de code Python nommée Streamlit (Adrien Treuille

et al., 2018). Streamlit permet de construire des applications web simples et interactives très rapidement sans avoir à gérer la partie interface et serveur séparément, le tout dans un seul et même langage de programmation. L'utilisation de Streamlit nous a permis de mettre en place, de modifier et d'affiner rapidement des démonstrations de nos outils basés sur l'IA.

4.7 Recherche ouverte et reproductibilité

La recherche ouverte est essentielle, notamment en IA pour faciliter l'adoption des outils, les améliorer et permettre leur analyse critique. Dans cette optique de science ouverte, nous avons utilisé différents outils pour partager nos travaux de recherche de façon libre et ouverte et les rendre reproductibles.

4.7.1 Développement open source et versionnage avec GitHub

Le code correspondant à l'ensemble des outils et des travaux présentés dans cette thèse est disponible de façon libre et open source sur mon profil GitHub (https://github.com/lambda-science). GitHub est une société fondée en 2008 qui permet le versionnage de code informatique. Le versionnage est un système qui permet de traquer les modifications apportées au code dans le temps et de garder une trace de toutes les versions qui ont existé. L'utilisation de tels systèmes facilite la collaboration en permettant à n'importe qui de contribuer au code, de proposer des améliorations ou de signaler des problèmes éventuels.

4.7.2 Développement de données et modèles IA open source avec HuggingFace

Si GitHub permet de versionner le code, il est aussi nécessaire d'avoir des outils permettant de rendre public et versionnable les modèles d'IA entrainés et les données utilisées. Mettre à disposition les modèles IA et les données utilisées de façon open source permet à la communauté d'évaluer de manière indépendante les modèles, mais aussi de les améliorer en entrainant de meilleurs à partir du jeu de données de base. Pour cela, nous avons mis en ligne les données utilisées pour entrainer le modèle SDH de **MyoQuant** ainsi que le modèle avec son code d'entrainement sur la plateforme HuggingFace à l'adresse : https://huggingface.co/corentinm7. HuggingFace est une société fondée en 2016 qui permet de mettre en libre accès des jeux de données et des modèles d'IA de manière open source. De plus, HuggingFace développe de nombreuses bibliothèques de code Python open source spécialisées en IA comme la bibliothèque nommée *transformers* qui permet l'entrainement de LLMs.

4.7.3 Suivi d'expérience avec *Weight and Biais*

L'entrainement d'un modèle IA est un processus long qui passe par de nombreuses phases de recherche sur les algorithmes et les architectures les plus performantes, les paramètres optimaux ou le meilleur pré-traitement des données pour obtenir un modèle performant. Pour traquer les performances des divers modèles entrainés et les comparer, des outils existent à l'instar de GitHub et HuggingFace. Dans cette thèse j'ai utilisé l'outil nommé *Weight and Bias* (https://wandb.ai/), une société fondée en 2017, pour traquer les performances des divers entrainements des modèles de **MyoQuant** et **NLMyo**. Dans les chapitres suivants, un lien vers les résultats en ligne est fourni pour accéder à l'ensemble des informations enregistrées.

4.7.4 Environnement de développement reproductible

Le dernier élément concernant la reproductibilité des travaux après avoir fourni le code, les données et les modèles de manière open source reste de s'assurer que le code puisse s'exécuter correctement quel que soit l'environnement informatique. Pour cela, pour chaque outil, nous fournissons deux éléments. En premier nous fournissons un environnement virtuel Python avec Poetry (https://python-poetry.org/) qui permet de spécifier les versions exactes de bibliothèques de code python utilisées ainsi que les versions de leurs dépendances. De plus, nous construisons une image Docker (https://www.docker.com/) pour chaque outil ce qui permet non seulement de contrôler la version de Python utilisée, mais aussi le système d'exploitation et les versions de ses diverses bibliothèques de codes annexes. L'utilisation d'images Docker couplée à un environnement virtuel Python spécifique permettent de s'assurer que notre code fonctionne et est reproductible sur n'importe quel matériel informatique.

4.7.5 Archivage du code, des données et des résultats avec Zenodo

Zenodo (European Organization For Nuclear Research et OpenAIRE, 2013, https://zenodo.org/) est un outil développé par le Conseil européen pour la recherche nucléaire (CERN) en 2013, qui permet d'archiver des travaux de recherche, du code et des données et d'y associer un identifiant unique citable (*Digital Object Identifier*, DOI). Dans le cadre de cette thèse, nous avons créé une archive Zenodo qui contient l'ensemble des éléments de cette thèse, c'est-à-dire : le manuscrit, le code des outils, les données non confidentielles utilisées, les modèles entrainés et les figures et les résultats. Le lien vers cette archive est https://doi.org/10.5281/zenodo.8431346.

Troisième partie

CONTRIBUTIONS

Chapitre 5

IMPatienT : annotation et exploration de données multimodales de patients

Dans un premier temps, nous avons développé **IMPatienT** (fig. 5.1) une application web qui permet l'annotation semi-automatique de compte rendu et d'images de biopsies musculaires. Cette application web, développée en 2020-2021, c'est à dire avant la mise à disposition des LLMs, permet de créer un jeu de données structuré (tableau de patients) à partir de données non structurées (texte libre et images). Ainsi dans ce contexte-là, pour exploiter des données sous la forme de texte libre, il est nécessaire de procéder à son annotation manuelle afin de les structurer. L'objectif d'**IMPatienT** est de mettre à disposition une interface qui permet la numérisation de ces données non structurée, de les préparer pour l'application des méthodes de ML traditionnelles et de fournir des outils d'explorations automatiques.

Pour structurer ces données biomédicales, nous utilisons une approche basée sur le concept d'ontologie : un vocabulaire standardisé pour décrire de manière unifiée les observations réalisées. S'il existe déjà des ontologies pour nommer les maladies (ORDO), les observations cliniques (HPO), et des nomemclatures pour les gènes et mutations (HUGO, HGVS), il n'existe aucune ontologie à ce jour permettant de décrire les observations histopathologiques dans les biopsies musculaires. Pour cela, **IMPatienT** intègre un module permettant en premier lieu la création de vocabulaire standard, que nous avons utilisé pour créer un vocabulaire standard des observations histopathologiques. Ensuite à partir de ce vocabulaire standard nous avons développé un module pour numériser et détecter de manière semi-automatique les termes du vocabulaire standard dans les comptes rendus de biopsies. Pour ajouter à l'aspect multimodal, nous avons développé un module de segmentation d'images assisté par intelligence artificielle. Ce module permet de rapidement annoter des images de biopsies avec les termes du vocabulaire standard afin de créer un jeu données annoté qui permet l'entrainement d'IA de segmentation d'images automatique. Enfin, un dernier module de visualisation des données enregistrées permet d'explorer en temps réel les données de patient numérisées dans l'application web.

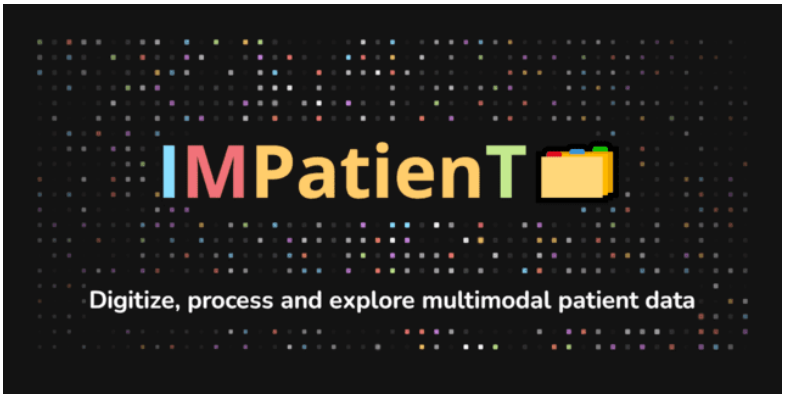


FIGURE 5.1 – **Logo d'IMPatienT**. IMPatienT est application web open-source disponible en version de démonstration en ligne. https://github.com/lambda-science/IMPatienT

5.1 Manuscrit

Le manuscrit d'**IMPatienT** a été soumis à la revue scientifique *"Journal of Neuromuscular Diseases"* et est présenté ci-dessous.

IMPatienT: an Integrated web application to digitize, process and explore Multimodal PATIENt daTa.

Corentin Meyer[a], Norma Beatriz Romero[b], Teresinha Evangelista[b], Brunot Cadot[c], Jocelyn Laporte[d], Anne Jeannin-Girardon[a], Pierre Collet[a], Kirsley Chennen[a], Olivier Poch[a*]

[a] *Complex Systems and Translational Bioinformatics (CSTB), ICube Laboratory, UMR 7357, University of Strasbourg, 1 rue Eugène Boeckel, 67000 Strasbourg, France.*

[b] *Neuromuscular Morphology Unit, Myology Institute, Reference Center of Neuromuscular Diseases Nord-Est-IDF, GHU Pitié-Salpêtrière, Paris, France*

[c] *Sorbonne Université, INSERM, Center for Research in Myology, Myology Institute, GHU Pitié-Salpêtrière, Paris, France*

[d] *Department Translational Medicine, IGBMC, CNRS UMR 7104, 1 rue Laurent Fries, 67404 Illkirch, France.*

* Correspondence to: Olivier Poch, CSTB – ICube UMR 7357, CRBS, 1 rue Eugène Boeckel, 67000 Strasbourg, Tel.: +33 3 68 85 32 95; Email: olivier.poch@unistra.fr

ABSTRACT

Medical acts, such as imaging, lead to the production of various medical text reports that describe the relevant findings. This induces multimodality in patient data by combining image data with free-text and consequently, multimodal data have become central to drive research and improve diagnoses. However, the exploitation of patient data is problematic as the ecosystem of analysis tools is fragmented according to the type of data (images, text, genetics), the task (processing, exploration) and domain of interest (clinical phenotype, histology). To address the challenges, we developed IMPatienT (**I**ntegrated digital **M**ultimodal **PATIEN**t da**T**a), a simple, flexible and open-source web application to digitize, process and explore multimodal patient data. IMPatienT has a modular architecture allowing to: (i) create a standard vocabulary for a domain, (ii) digitize and process free-text data, (iii) annotate images and perform image segmentation, (iv) generate a visualization dashboard and provide diagnosis decision support. To demonstrate the advantages of IMPatienT, we present a use case on a corpus of 40 simulated muscle biopsy reports of congenital myopathy patients. As IMPatienT provides users with the ability to design their own vocabulary, it can be adapted to any research domain and can be used as a patient registry for exploratory data analysis. A demo instance of the application is available at https://impatient.lbgi.fr/.

Keywords: Muscular Diseases; Histology; Image Processing, Computer-assisted; Diagnosis, Computer-assisted; Electronic Health Records; Artificial Intelligence

INTRODUCTION

Patient data now incorporates the results of numerous technologies, including imaging, next-generation sequencing and more recently wearable devices. Furthermore, medical acts such as echography, radiology or histology, produce imaging data that are generally combined with medical reports to describe the relevant findings. Thus, multimodality is induced in patient data, as imaging data is inherently linked to free-text reports. The link between image and report data is crucial as raw images can be re-interpreted during the patient's medical journey with new domain knowledge or by different experts leading to complementary reports. The use of multimodal data has been shown to increase disease understanding and diagnosis [1–4]. For example, Venugopalan *et al.* integrated genetic data with image data and medical records (free-text data) to improve diagnosis of Alzheimer's disease [4]. In Mendelian diseases, integration of multiple levels of information is key to the establishment of a diagnosis. For instance, in congenital myopathies (CM), a combination of muscle biopsy analysis (imaging information) with medical records and sequencing data is essential for differential diagnosis between CM subtypes [5–7].

Centralization of multimodal data using dedicated software is essential to implement such an approach. First, multimodal patient data needs to be processed in an integrated way to preserve this link in a single database or data warehouse. Second, useful tools to process and explore multimodal data are essential to drive research and improve diagnosis.

Unfortunately, the ecosystem of software tools for the exploitation of patient data is heavily fragmented, according to the type of data (images, text, genetic sequences), the task to be performed (digitization, processing, exploration) and the domain of interest (clinical phenotype, histology, etc.). Exploitation tools can be divided into two main categories: (i) software to process the data and (ii) software to explore the data.

Clinical reports are generally written using free-text, and therefore processing relies on the use of a standard vocabulary, such as the Unified Medical Language System (UMLS) [8] or the Human Phenotype Ontology (HPO)[9]. Several tools have been developed to easily manage and extend these standard vocabularies, including Protégé [10]. Text mining processes have been developed that exploit these standard vocabularies to automatically detect important keywords in free-text data. For example, Doc2HPO [11] can extract a list of HPO terms from free-text medical records. Other software packages, *e.g.* Phenotips [12], have been developed to centralize and process general patient information, including demographics, pedigree, common measurements, phenotypes and genetic results. SAMS [13] and RD-Connect PhenoStore [14] are further examples of web applications that aim to perform deep phenotyping of patients by building a single database of standardized patient data using well-established ontologies such as HPO.

A number of tools have been developed to analyze and explore patient data, based on a list of HPO terms describing a patient's specific phenotypic profile. For example, Phenolyzer [15] and Phenomizer [16] can be used to help prioritize candidate genes or rank the best-matching diseases. However, these tools are restricted to the use of HPO terms to describe the patient's profile and are not compatible with other ontologies. Ontology agnostic algorithms have also been developed that predict an outcome based on a list of terms from any normalized vocabulary, such as the Bayesian Ontology Query Algorithm (BOQA) [17].

Finally, for imaging data, software to process and annotate gigapixel scale microscopy images are widely used, including Cytomine [18], SlideRunner [19] and Ilastik [20]. While Cytomine incorporates an ontology builder and complex image processing tools, it is restricted to image data only. For exploitation of patient images, guidelines and frameworks have been proposed to standardize the measurement of pathological features for example from DICOM lung images [21]. Some multimodal approaches such as ClinPhen [22] and Exomiser [23] have

successfully combined multiple levels of information with both phenotype information (HPO terms) and genetic information (variants) to rank candidate genes in Mendelian diseases. Other tools such as INTEGRO [24] have been developed to automatically mine disease-gene associations for a specific input disease from multiple curated sources of knowledge.

This large ecosystem of tools highlights the need for an integrated tool that can: (i) process and explore patient data, (ii) manage multimodal data (text and images), and (iii) work in any domain of interest.

In this study, we present IMPatienT (**I**ntegrated digital **M**ultimodal **PATIEN**t da**T**a), a free and open-source web application designed to provide an integrated tool to digitize, process and explore multimodal patient data**.** IMPatienT is a turnkey solution that can aggregate patient data and provides simple tools and interfaces allowing clinicians to easily extract information from multimodal patient data. IMPatienT is based on a modular architecture, and currently incorporates four components to: (i) create a standard vocabulary describing a domain of interest, (ii) digitize and process free-text records by automatically mapping them to a set of standard terms, (iii) annotate and segment images with standard vocabulary, and (iv) generate a dashboard with automatic visualizations to explore the patient data and perform automatic diagnosis suggestions.

We demonstrate the usefulness of IMPatienT on a set of congenital myopathy (CM) cases. CM are a family of rare genetic diseases, including multiple distinct subtypes, that still lack proper diagnosis with more than 50% of patients without a genetic cause identified [25]. We exploited IMPatienT to create a vocabulary of standard muscle-histology terms that were then used to process patient histological records and annotate biopsy images. Finally, multiple exploratory visualizations were automatically generated.

MATERIALS AND METHODS

IMPatienT is a web application developed with the Flask micro-framework, which is a Python-based web framework. Figure 1 illustrates the global organization of the web application, currently composed of four modules: (i) Standard Vocabulary Creator, (ii) Report Digitization, (iii) Image Annotation, and (iv) Automatic Visualization Dashboard. All modules incorporate free, open-source and well-maintained libraries that are described in detail in the corresponding sections.

Module 1: Standard Vocabulary Creator

The standard vocabulary creator module allows to create and modify a hierarchical list of vocabulary terms with rich definitions that can be used as image annotation classes, for processing of text reports, or diagnosis decision support. The standard vocabulary creator is an essential module as it interacts with all subsequent modules.

Figure 2 shows a screenshot of the page used to create and manage the standard vocabulary tree. The tree is generated and rendered with the JavaScript library JSTree (version 3.3.12). Each node in the tree represents a vocabulary term, and each term can have only one parent. The ergonomic drag and drop system uses the graphical user interface (GUI) and allows the user to intuitively and quickly edit and reorganize the vocabulary by adding new terms or modifying existing ones. For each created node (vocabulary term), the user can assign a name and organize the tree structure (hierarchy).

Each term in the tree is associated with nine optional properties, available *via* the vocabulary term (node) detailed form. Four properties are defined by the user: description, list of synonyms, translation in another language, and use as annotation class. Two properties are automatically generated: the unique identifier (ID) and the hexadecimal color associated with the term (for image annotation). Additional term properties, including associated diagnosis/disease class, associated genes, and the list of positively correlated terms (*i.e.* co-occurring terms in reports), are extracted from patient records registered in the database.

Finally, if the user defines an alternative translation for terms, there is an “invert vocabulary language” button to conveniently switch between languages. For instance, the user can create a vocabulary in any language and define the translation in English, then switch between the two display modes easily.

Module 2: Report Digitization

The standard vocabulary terms are used to process documents that are in a free-text format. Module 2 uses a semi-automatic approach for digitization and processing of free-text reports that combines fast automatic detection of terms with manual reviewing of the detection. The interface of Module 2 is a form divided into four parts (Figure 3).

In the first part of the digitization form (Figure 3a), a PDF file of the free-text report can be uploaded for natural language processing (NLP) of the content. The text of the PDF report is automatically extracted and processed with NLP. The NLP method is only used to detect histological terms defined in the standard vocabulary. Detected standard vocabulary terms are highlighted (see corresponding section below “Optical Character Recognition and

Vocabulary Term Detection"). Highlighted terms allow to easily identify which standard vocabulary terms were detected as present in positive or in negative form. This is useful for quantitative performance assessment.

The second part (Figure 3b) of the digitization form contains patient information, such as patient ID, document ID, patient age. This section also allows the user to input patient information that is not defined by the standard vocabulary and thus, not processed in the NLP section. For example, IMPatienT exploits well-established ontologies to normalize the genetic diagnosis and phenotypes (Figure 4). For example, in the gene field, when the user inputs a character string, gene symbols are retrieved from the HUGO Gene Nomenclature Committee (HGNC) and suggested [26]. Mutation notations are formatted according to the Human Genome Variation Society (HGVS) sequence variant nomenclature [27]. Phenotypes are retrieved and suggested using the HPO. These fields do not contain patient-identifying data and are optional.

The third part of the digitization form (Figure 3c) contains the standard vocabulary tree viewer with an absence/presence slider. This section allows the user to correct the automatic detection of the NLP method or to add new observations. Each vocabulary term can be marked as present, absent or no information. For terms marked as present, the slider is used to indicate a notion of quantity or certainty of the term. For example, the statement "There are a small number of fibers containing rods" can be annotated by hand by setting the vocabulary "Rods" to the value "Present" with a low quantity value. For terms that have been automatically detected, this slider value is automatically set to 0 (present in a negated sentence) or 1 (present).

The fourth part (Fig 3d) of the form allows the user to input comments and a final diagnosis for the patient. Disease names are extracted and suggested from the Orphanet [28] knowledge base. It also includes an automatic diagnosis suggestion based on already registered patients and the BOQA algorithm [17] (see the corresponding section below "Method for Patient Disease Suggestions").

Optical Character Recognition and Vocabulary Term Detection

The patient report digitization in module 2 is facilitated by an automatic text recognition and keyword detection method. The user uploads a PDF version of the text reports to perform Optical Character Recognition (OCR), followed by Natural Language Processing (NLP) to automatically detect terms from the standard vocabulary in the report. The NLP method only matches the raw text to the standard vocabulary defined in Standard Vocabulary Module 1. Figure 5 shows the workflow of the vocabulary terms detection method. First, the PDF file is converted to plain text using the Tesseract OCR (implemented in python as pyTesseract). Then, the text is processed with Spacy, an NLP python library, by splitting the text into sentences and then into individual words. The resulting list

of sentences is processed to detect negation using a simple implementation of the concept of NegEx [29]. An *n*-gram (monograms, digrams, and trigrams) procedure is applied to the list of words to identify contiguous words in the context of all the sentences of the report. The *n*-grams are then mapped against the user-created standard vocabulary using fuzzy partial matching (based on Levenshtein distance) with a score threshold of 0.8. Matched keywords are kept and shown on the interface by green or red highlighting of the detected text using the Mark.JS JavaScript library (green indicates the presence of the keyword, red indicates the presence in a negated sentence). Keywords are also automatically marked as present or absent (negated) in the vocabulary tree.

Disease Suggestions

The report digitization module 2 contains a disease recommendation algorithm inspired by the BOQA algorithm described by Bauer *et al.* [17]. Basically, the algorithm computes the similarity between a list of input vocabulary terms annotated as "present" for a patient (the query) and a simulated patient profile for each disease class (model report) that is generated based on the data from already registered patients.

We implemented this algorithm in python, and modified it to use the frequencies of vocabulary terms per disease for the generation of the model report instead of the initial deterministic way (not frequency aware). This means that the model report is generated based on the probability (frequency) of each vocabulary term. For example, if disease A is annotated with vocabulary term B at a frequency=0.9 and vocabulary term C at a frequency=0.1, the generated model report for disease A will have a probability=0.9 of containing vocabulary term B and a probability=0.1 of containing vocabulary term C.

Due to the stochastic nature of the generation of the model report, for any given prediction, the generation and computation of the similarity with the query is repeated 50 times. For each repetition, if a disease has a prediction probability>0.5, it is considered to be the best prediction, otherwise the prediction is "no prediction". Finally, of the 50 repetitions, the prediction with the highest occurrence is taken as the final prediction.

Module 3: AI-Assisted Image Annotation Using Automatic Segmentation

To process patient image data, we developed the image annotation module (module 3) to upload, annotate and perform image segmentation with standard vocabulary terms. This module is based on the *"interactive image segmentation with Dash and Scikit-image"* demonstration application [30–32]. The original source code was modified to be compatible with the standard vocabulary tree and the database.

The interactive interface to annotate image features with standard vocabulary terms is presented in figures 6a and 6b. The interface allows the user to draw a free-shape area (annotation) associated with a standard vocabulary term (class). Then, with a minimal number of user annotations, the whole image is segmented based on the annotations (shapes) provided by the user.

To perform image segmentation, on the server side, local features (intensity, edges, texture) are extracted from the labeled areas of the image and are used to train a dedicated AI random-forest classifier model. This dedicated model is then applied to predict similar areas in the whole image. Finally, every pixel of the image is labeled with a standard vocabulary term corresponding to the AI prediction based on the annotations.

The segmentation is entirely interactive. After the initial segmentation, the user can correct the classification by adding more annotation shapes to the image and can modify the paintbrush width setting to make more precise annotation marks. In addition, the stringency range parameter of the model can be adapted using the slider to modify the model behavior and automatically recompute the segmentation in real time.

Results of the segmentation are retrievable as a single archive including the raw image, the annotations (JSON format), the random-forest trained classifier, the blended image and the segmentation mask image.

Module 4: Automatic Visualization Dashboard

The automatic visualization dashboard module is designed to perform exploratory data analysis by generating multiple graphs based on the patient data in the database. All visualizations are created using Plotly, a python graph library, that allows the creation of interactive graphs.

Interaction Between the Modules

IMPatienT is divided into four modules that are interconnected. The standard vocabulary module provides the vocabulary used for the image annotation module and for the NLP method used for the (histologic) standard vocabulary term detection in the report digitization module. Any modification in the vocabulary is automatically propagated to these modules, updating the form templates and triggering the recalculation of all visualizations with the latest vocabulary information. Any modification to the standard vocabulary also updates all patients in the database to the latest version of the vocabulary, meaning that term names and definitions will be updated, and deleted terms will be marked as outdated. Adding patient information in the database, whether they are text reports (module 2) or image data (module 3), will automatically update the visualization dashboard with the latest patient information in the database. The term frequency statistics calculated by the visualization dashboard and used by

the disease suggestion algorithm are automatically updated as well, increasing live performances. The visualization dashboard is also directly linked to the standard vocabulary and during the generation of the visualizations, the rich definition of the standard terms is updated with newly associated genes, diagnosis and positively correlated terms.

Application Security and Personal Data

IMPatienT is developed as a free and open-source project meaning that the code can be audited by anyone in the GitHub code repository (https://github.com/lambda-science/IMPatienT).

The code is regularly scanned for known issues and outdated libraries to mitigate security issues. There is no patient-identifying data kept in the database, only a custom identifier and age. The synthetic dataset generated and analyzed during the current study is also available in the same repository. No name or date of birth is required or stored. Additionally, access to all modules and data entered via the web application is restricted by a login-page and user accounts can only be created by the administrator of the platform. No user information is stored except for the username, email and salted and hashed passwords.

RESULTS

IMPatienT is an interactive and user-friendly web application that integrates a semi-automatic approach for text and image data digitization, processing, and exploration. Due to its modular architecture and its standard vocabulary creator, it has a wide range of potential uses.

IMPatienT Main Functionalities

Table 1 shows the main functionalities of IMPatienT compared to other similar tools used in the community. IMPatienT integrates tools that are simple, portable, easy to implement and similar to multiple state-of-the-art solutions but in a single platform. Out of 18 selected features, IMPatienT integrates 14 of them versus a mean of 4.4 for other software with the best ones being SAMS and PhenoStore integrating 6 features each. Nevertheless, software such as SAMS, PhenoStore, Phenotips and Cytomine each integrate features that are not yet present in IMPatienT.

IMPatienT implements novel functionalities to process and exploit patient data. For example, IMPatienT is compatible with any research domain thanks to its standard vocabulary builder. Also, with the OCR/NLP method, IMPatienT can process histologic text reports, allowing the user to exploit scanned documents. Finally, IMPatienT

provides useful tools to exploit patient data with the various visualizations, the term, frequency table, correlation matrix and the automatic enrichment of the vocabulary term definitions (associated genes and diseases).

IMPatienT Usage

Figure 1 shows how the user can interact with the web application to digitize, process, and explore patient data. In IMPatienT, modules can be used independently, allowing users to only use the tools they need. For example, a user might only have text report data, in this case they would be able to use the standard vocabulary creator, the report digitization tools and the visualization dashboard to process and explore their data. In another scenario, a user could only be interested in annotating an image dataset using a shared standard vocabulary that can be modified and updated collaboratively. In this use case, they would be able to only use the standard vocabulary creator and the image annotation module. However, the main strength of IMPatienT lies in the multimodal approach it provides and the strong interactions between modules.

For the complete multimodal approach, the first step is to create a standard vocabulary using the Standard Vocabulary Creator interface (module 1). The user only needs to create a few terms (nodes) to begin using the web application. Defining the properties of the terms (definition, synonyms, etc.) is optional, and organizing them in a hierarchical structure is also optional.

In the second step, the user can start digitizing patient reports using module 2. This can be done either manually by filling out the form in module 2 and checking terms as present or absent in a given report, or automatically using the Vocabulary Term Matching method to process a PDF version of the report. Using module 3, the user can also upload, annotate, and segment image data.

Finally, the user can explore multiple visualizations (histograms, correlation matrix, confusion matrix, frequency tables) that are automatically generated in module 4. All data entered *via* the web application are retrievable in standard formats, including the whole database of reports as a single SQLite3 file or CSV files, the images and their segmentation models and masks as a GZIP archive, the standard vocabulary with annotation as a JSON file and various graphs and tables as JSON or PNG files.

Use Case: Congenital Myopathy Histology Reports

As a use case of IMPatienT, we focused on congenital myopathies (CM). We used the standard vocabulary creator to create a sample muscle histology standard vocabulary based on common terms used in muscle biopsy reports from the Paris Institute of Myology. Then, we inserted 40 digitally generated patients in the database with random

sampling of standard vocabulary terms and associated a gene and disease class from a list of common CM genes and three recurring CM subtypes (nemaline myopathy, core myopathy and centronuclear myopathy). All these data are available on the demo instance of IMPatienT (https://impatient.lbgi.fr/).

For text data, Supplementary Figure S1 shows the results of the automatic NLP method applied to an artificial muscle histology report. Twenty-two keywords were detected that match to the standard vocabulary and seven of them were detected in negated sentences (red highlight). Out of the twenty-two keywords, eighteen were correctly detected and one was detected in the wrong state of negation: "abnormal fiber differentiation" is highlighted as negated although it is present is a non-negated sentence part. Three keywords (fiber type, internalized nuclei, centralized nuclei) were detected as matching for multiple keywords from the vocabulary due to high similarity. For example, the keywords "internalized nuclei" and "centralized nuclei" have a similarity score of 86 using the Levenstein distance. Two keywords defined in the standard vocabulary were missed and not highlighted: "biopsy looks abnormal" ("abnormal biopsy" in the vocabulary) and "purplish shade" ("purplish aspect" in the vocabulary).

For the image data, Figure 7 shows an example of the segmentation of a biopsy image, where we annotated the cytoplasm of the cells (green), intercellular spaces (black) and cell nuclei (red). The raw image (Figure 7a) is annotated with free-shape areas associated with standard vocabulary terms (Figure 7b). Then, the whole image is automatically segmented based on the annotations, producing the segmentation mask where each pixel is associated with a class (Figures 7c, 7d).

The automatic visualization dashboard was used to generate the six visualizations provided in Figure 8. These visualizations include a breakdown of the patients in the database by age, genes, or diagnosis (Figure 8a). A correlation matrix (using Pearson correlation coefficient) between the occurrences of standard vocabulary terms is generated (Fig 8b), which can serve as a starting point for exploration of co-occurrence of features in patients. The confusion matrix of the final diagnosis of patients versus the suggested diagnosis with BOQA (Fig 8c) allows the user to monitor the accuracy of the disease suggestion function. In addition, a histogram showing the classification of patients without a final diagnosis is provided to indicate possible prognosis of undiagnosed patients (Figure 8d). Finally, the frequency of each standard vocabulary term by gene and by disease is automatically calculated and shown in two tables (Supplementary Tables S1 and S2).

DISCUSSION

IMPatienT is a platform that simplifies the digitization, processing, and exploration of both textual and image patient data. The web application is centered around the concept of a standard vocabulary tree that is easy to create and used to process text and image data. This allows IMPatienT to work with patient data from domains that still lack a consensus ontology and rely on well-established ontologies for patient data, such as HPO for phenotypes, Orphanet for disease names or HGCN/HGVS for genetic diagnoses.

The semi-automatic approach implemented in IMPatienT offers faster digitization processes while ensuring accuracy through manual review. This is achieved by analyzing text data using OCR and NLP to automatically match the text to the standard vocabulary, followed by manual correction. For image data, the user first provides sparse annotations on the image, which are then used to compute an automatic segmentation of the whole image. For data exploration, IMPatienT uses a fully automatic approach including various visualizations as well as diagnosis suggestions, while allowing the user to extract the processed data in a standard format for further analysis (database, images, frequency tables).

IMPatienT aims to integrate multiple approaches in a unified platform with two main objectives: universality (*i.e.* not restricted to a specific domain) and multimodality (*i.e.* integration of multiple data types). To our knowledge, other tools similar to IMPatienT do not fulfill both objectives.

We performed a comparison of the main functionalities of IMPatienT with other tools used in the community. Phenotips, SAMS and PhenoStore are similar to IMPatienT as they are designed as a patient information database. However, they are restricted to processing patient phenotype data using HPO and do not integrate multimodal data. IMPatienT goes further by allowing for custom observations with the vocabulary builder and with automatic digitization with OCR/NLP as well as integrating tools to exploit image data.

Other tools are available that implement the functionality of one or two IMPatienT modules. For example, Doc2HPO is a tool that also uses a semi-automatic approach to digitize clinical text according to a list of HPO terms, based on NLP methods and negation detection. However, as Doc2HPO is restricted to HPO, it does not provide custom vocabulary tree facilities. In contrast IMPatienT is suitable for digitization of text data from any domain of interest.

For image data, software such as Cytomine and Ilastik are widely used and perform well on biological data, but they do not allow the user to take into consideration the multimodal aspects of patient data by keeping the raw image and the expert interpretation (histological report) in a single database along with a collaborative and rich user-defined ontology.

Finally, in IMPatienT we reimplemented the diagnosis support algorithm called BOQA that is also used in Phenomizer, a tool to rank a list of the top matching diseases based on a list of input HPO terms. We modified the algorithm to consider frequencies of terms by disease to have meaningful predictions. In contrast, BOQA uses binary states for terms (terms are marked as present or absent) and is not compatible with numeric features. In the future, it will be interesting to implement a more complex system such as explainable AI with learning classifier systems [33]. This should improve accuracy, explainability, and handling of quantitative values, although at the cost of computational power.

IMPatienT still lacks some features compared to other tools, such as a pedigree editor, support for DICOM and gigapixel images and phenotypic data export to the Phenopacket format. In the future, we plan to further develop IMPatienT by adding these features to the interface. We also want to explore the automatization of the standard vocabulary creation with the analysis of a complete corpus of text. For text analysis, we intend to implement additional context comprehension, *i.e.* not only negation but also hypothetical statements, uncertainty and family context as well as better text-vocabulary terms matching. Finally, we plan to expand the scope of the OCR/NLP method by integrating existing NLP tools to automatically detect HPO terms, gene symbols and disease names in the report text.

Acknowledgements

We thank the BiGEst-ICube platform for their assistance. We thank the Agence Nationale de la Recherche (ANR), 80 | Prime CNRS (MYO-xIA Project), the University of Strasbourg and INSERM for funding this work.

Conflicts of Interest

The authors have no conflict of interest to report.

References

[1] Kerr WT, Hwang ES, Raman KR, Barritt SE, Patel AB, Le JM, et al. Multimodal diagnosis of epilepsy using conditional dependence and multiple imputation. 2014 Int. Workshop Pattern Recognit. Neuroimaging, 2014, p. 1–4. https://doi.org/10.1109/PRNI.2014.6858526.

[2] Yan R, Ren F, Rao X, Shi B, Xiang T, Zhang L, et al. Integration of Multimodal Data for Breast Cancer Classification Using a Hybrid Deep Learning Method. In: Huang D-S, Bevilacqua V, Premaratne P, editors. Intell. Comput. Theor. Appl., Cham: Springer International Publishing; 2019, p. 460–9. https://doi.org/10.1007/978-3-030-26763-6_44.

[3] Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. Transl Res J Lab Clin Med 2018;194:56–67. https://doi.org/10.1016/j.trsl.2018.01.001.

[4] Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 2021;11:3254. https://doi.org/10.1038/s41598-020-74399-w.

[5] North KN, Wang CH, Clarke N, Jungbluth H, Vainzof M, Dowling JJ, et al. Approach to the diagnosis of congenital myopathies. Neuromuscul Disord NMD 2014;24:97–116. https://doi.org/10.1016/j.nmd.2013.11.003.

[6] Cassandrini D, Trovato R, Rubegni A, Lenzi S, Fiorillo C, Baldacci J, et al. Congenital myopathies: clinical phenotypes and new diagnostic tools. Ital J Pediatr 2017;43:101. https://doi.org/10.1186/s13052-017-0419-z.

[7] Böhm J, Vasli N, Malfatti E, Le Gras S, Feger C, Jost B, et al. An integrated diagnosis strategy for congenital myopathies. PloS One 2013;8:e67527. https://doi.org/10.1371/journal.pone.0067527.

[8] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267–70. https://doi.org/10.1093/nar/gkh061.

[9] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res 2021;49:D1207–17. https://doi.org/10.1093/nar/gkaa1043.

[10] Musen MA. The Protégé Project: A Look Back and a Look Forward. AI Matters 2015;1:4–12. https://doi.org/10.1145/2757001.2757003.

[11] Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res 2019;47:W566–70. https://doi.org/10.1093/nar/gkz386.

[12] Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Hum Mutat 2013;34:1057–65. https://doi.org/10.1002/humu.22347.

[13] Steinhaus R, Proft S, Seelow E, Schalau T, Robinson PN, Seelow D. Deep phenotyping: symptom annotation made simple with SAMS. Nucleic Acids Res 2022:gkac329. https://doi.org/10.1093/nar/gkac329.

[14] Laurie S, Piscia D, Matalonga L, Corvó A, Fernández-Callejo M, Garcia-Linares C, et al. The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. Hum Mutat 2022;43:717–33. https://doi.org/10.1002/humu.24353.

[15] Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods 2015;12:841–3. https://doi.org/10.1038/nmeth.3484.

[16] Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet 2009;85:457–64. https://doi.org/10.1016/j.ajhg.2009.09.003.

[17] Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics 2012;28:2502–8. https://doi.org/10.1093/bioinformatics/bts471.

[18] Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics 2016;32:1395–401. https://doi.org/10.1093/bioinformatics/btw013.

[19] Aubreville M, Bertram C, Klopfleisch R, Maier A. SlideRunner - A Tool for Massive Cell Annotations in Whole Slide Images. ArXiv180202347 Cs 2018:309–14. https://doi.org/10.1007/978-3-662-56537-7_81.

[20] Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, et al. ilastik: interactive machine learning for (bio)image analysis. Nat Methods 2019;16:1226–32. https://doi.org/10.1038/s41592-019-0582-9.

[21] Cinaglia P, Tradigo G, Cascini GL, Zumpano E, Veltri P. A framework for the decomposition and features extraction from lung DICOM images. Proc. 22nd Int. Database Eng. Appl. Symp., New York, NY, USA: Association for Computing Machinery; 2018, p. 31–6. https://doi.org/10.1145/3216122.3216127.

[22] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med 2019;21:1585–93. https://doi.org/10.1038/s41436-018-0381-1.

[23] Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc 2015;10:2004–15. https://doi.org/10.1038/nprot.2015.124.

[24] Cinaglia P, Guzzi PH, Veltri P. INTEGRO: an algorithm for data-integration and disease-gene association. 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM, 2018, p. 2076–81. https://doi.org/10.1109/BIBM.2018.8621193.

[25] Jungbluth H, Treves S, Zorzato F, Sarkozy A, Ochala J, Sewry C, et al. Congenital myopathies: disorders of excitation-contraction coupling and muscle contraction. Nat Rev Neurol 2018;14:151–67. https://doi.org/10.1038/nrneurol.2017.191.

[26] Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res 2021;49:D939–46. https://doi.org/10.1093/nar/gkaa980.

[27] den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat 2016;37:564–9. https://doi.org/10.1002/humu.22981.

[28] INSERM. Orphanet: an online database of rare diseases and orphan drugs 1997. http://www.orpha.net (accessed February 13, 2022).

[29] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform 2001;34:301–10. https://doi.org/10.1006/jbin.2001.1029.

[30] Gouillart E. Interactive Machine Learning - Image segmentation. GitHub 2020. https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation (accessed November 23, 2021).

[31] Walt S van der, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ 2014;2:e453. https://doi.org/10.7717/peerj.453.

[32] Hossain S. Visualization of Bioinformatics Data with Dash Bio. Proc 18th Python Sci Conf 2019:126–33. https://doi.org/10.25080/Majora-7ddc1dd1-012.

[33] Urbanowicz RJ, Moore JH. ExSTraCS 2.0: Description and Evaluation of a Scalable Learning Classifier System. Evol Intell 2015;8:89–116. https://doi.org/10.1007/s12065-015-0128-8.

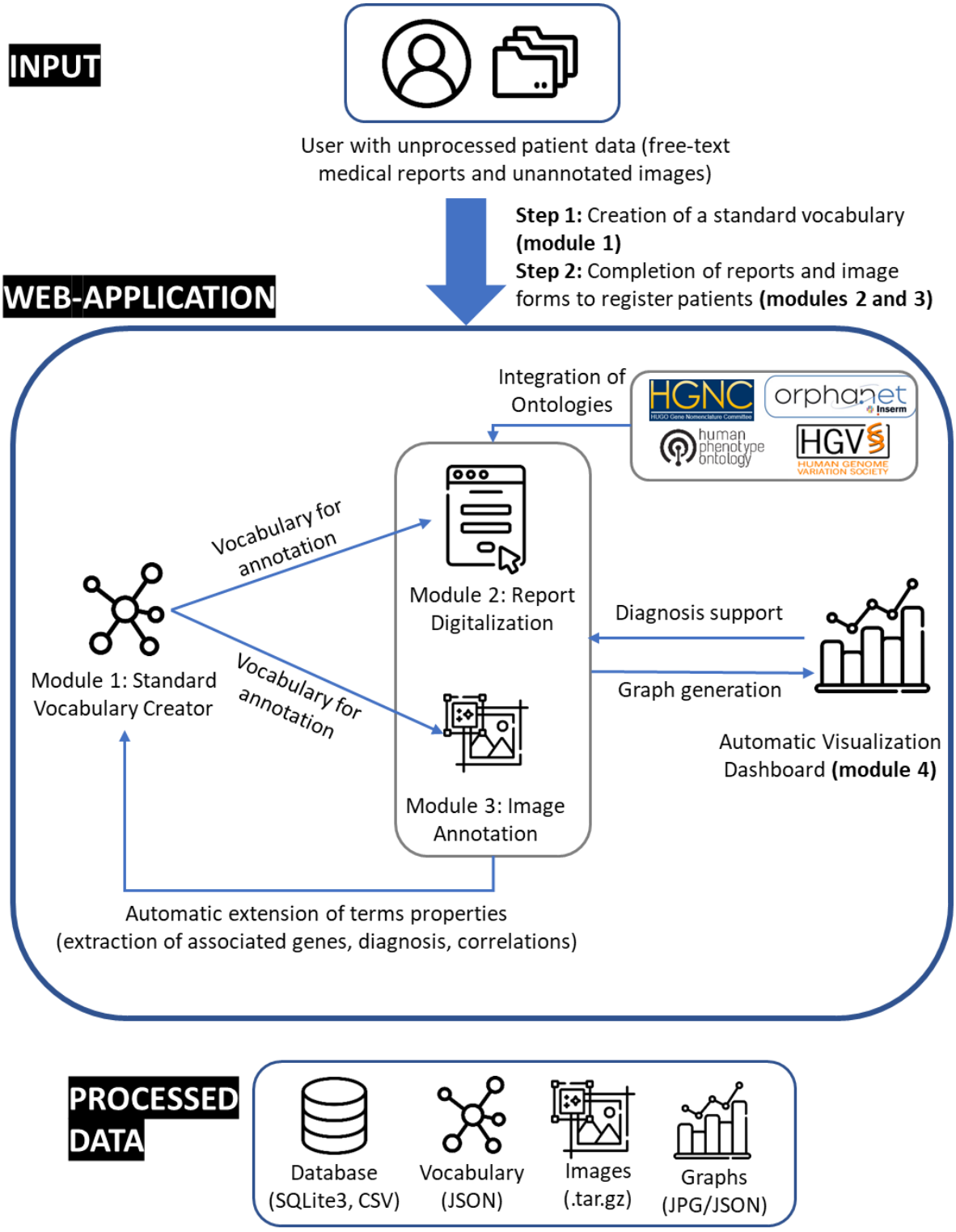


Figure 1: Organization of IMPatienT web application

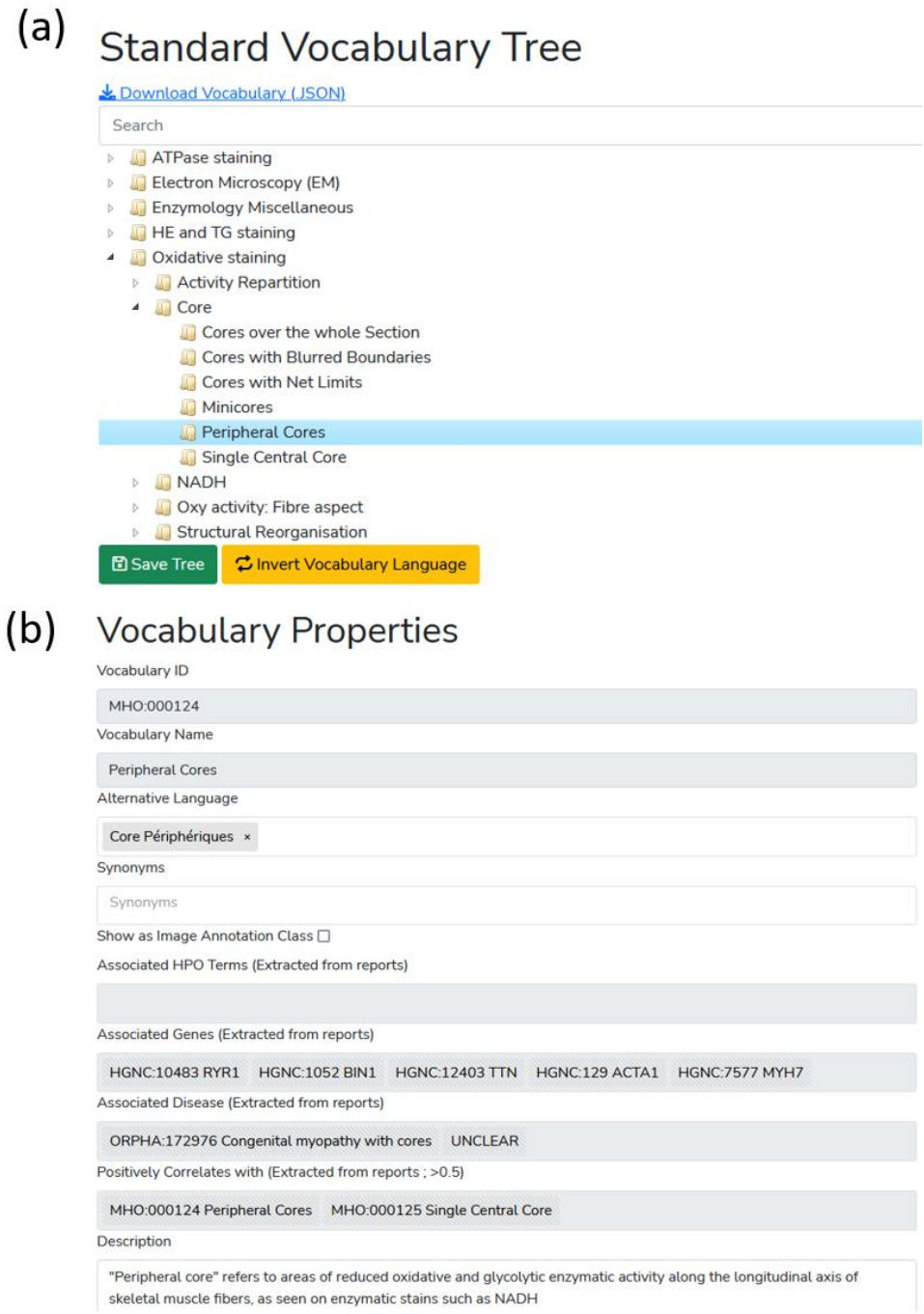


Figure 2: Screenshot of the Standard Vocabulary Creator module (module 1). (a) The hierarchical structure viewer and editor tool that supports drag and drop modification and creation/deletion/modification using the mouse. **(b)** The properties of the selected term node with its unique identifier (ID), name, alternative language translation, synonyms, description, associated genes and diseases and correlated terms extracted from the application instance database.

Figure 3: Screenshot of the interface for the report digitalization module. (a) PDF upload section for automatic keyword detection in the text. Detected keywords have a green background, detected and negated keywords have a red background. **(b)** Patient information section (age, document ID, gene, mutation, phenotype). **(c)** Standard vocabulary tree viewer to select keywords with associated slider to manually indicate keyword value (absence or presence level). Keywords marked as present are indicated with a green check mark, absent keywords are marked with a red cross. **(d)** Overview section of all annotated terms, diagnosis selection and commentaries with automatic diagnosis support using the BOQA algorithm.

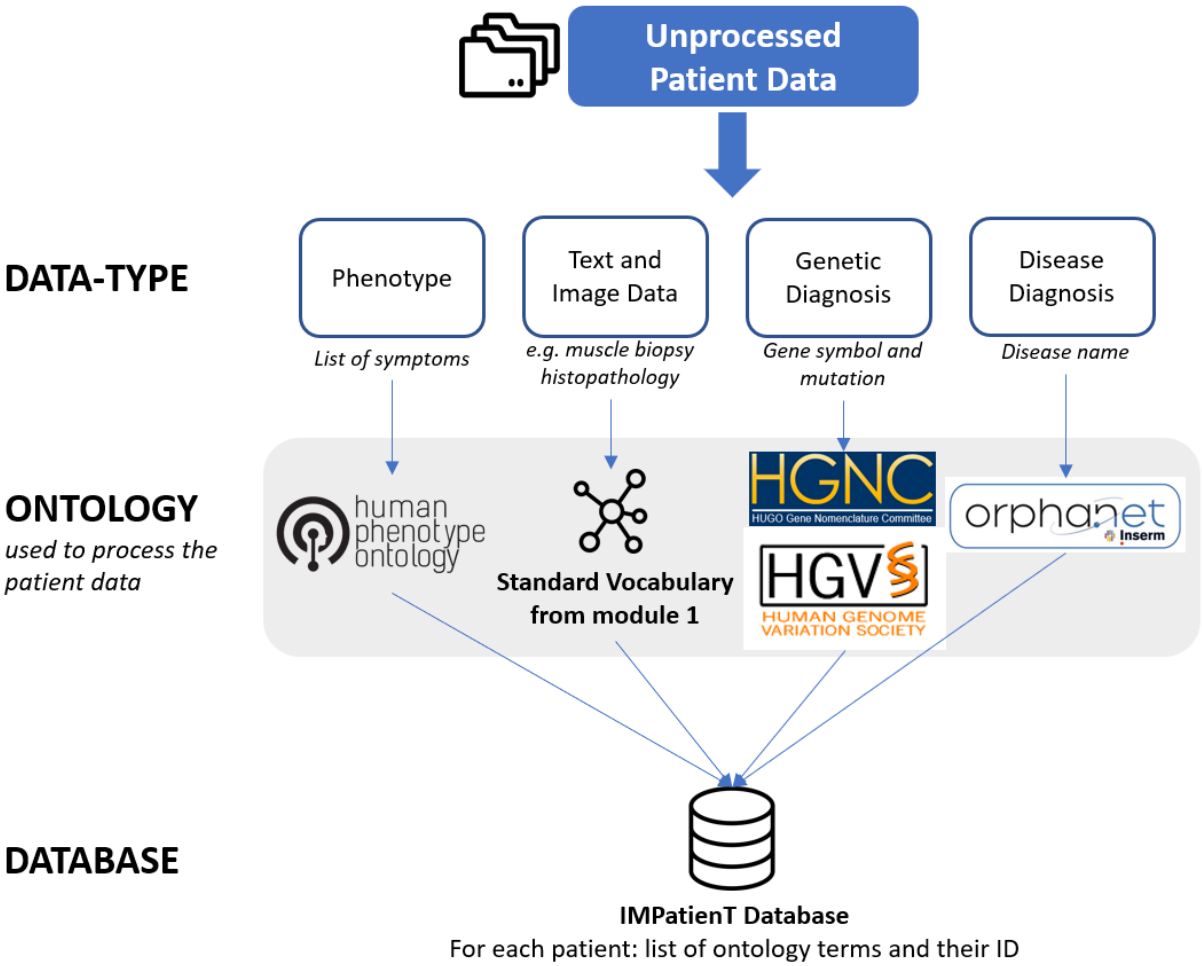


Figure 4: Overview of the ontologies used by IMPatienT to process patient data in the report digitization module (module 2).

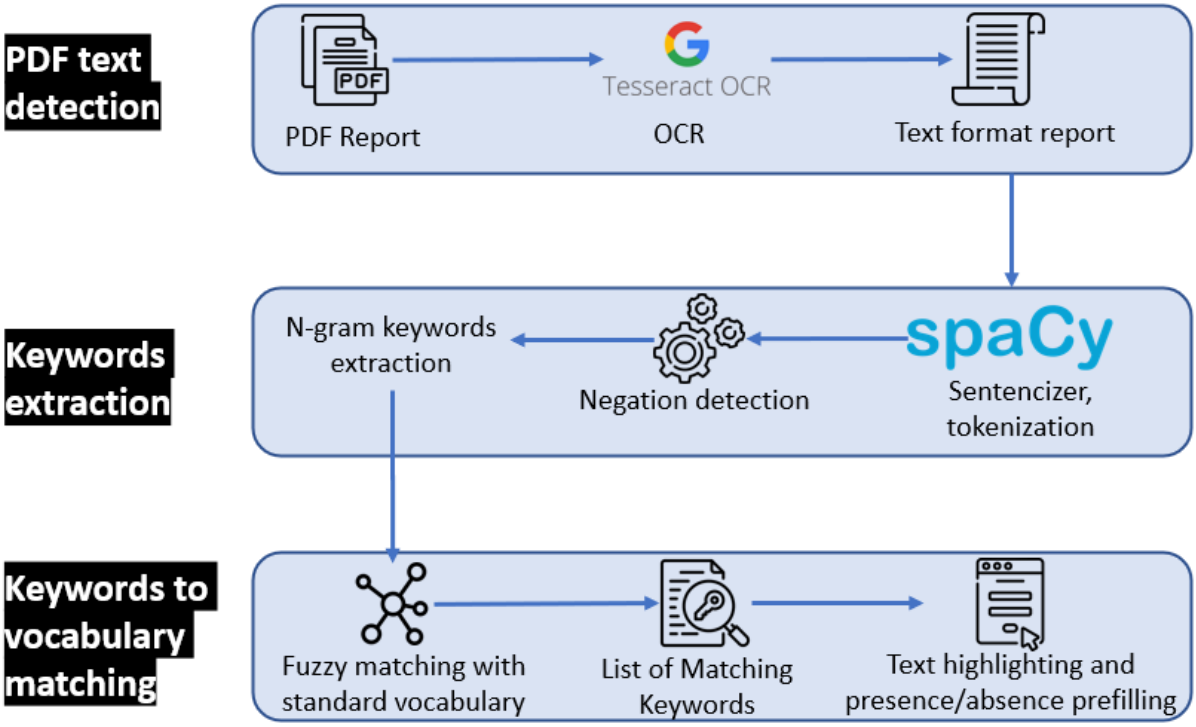


Figure 5: Optical character recognition and vocabulary term detection method used in the report digitization module (module 2) to automatically analyze free-text reports.

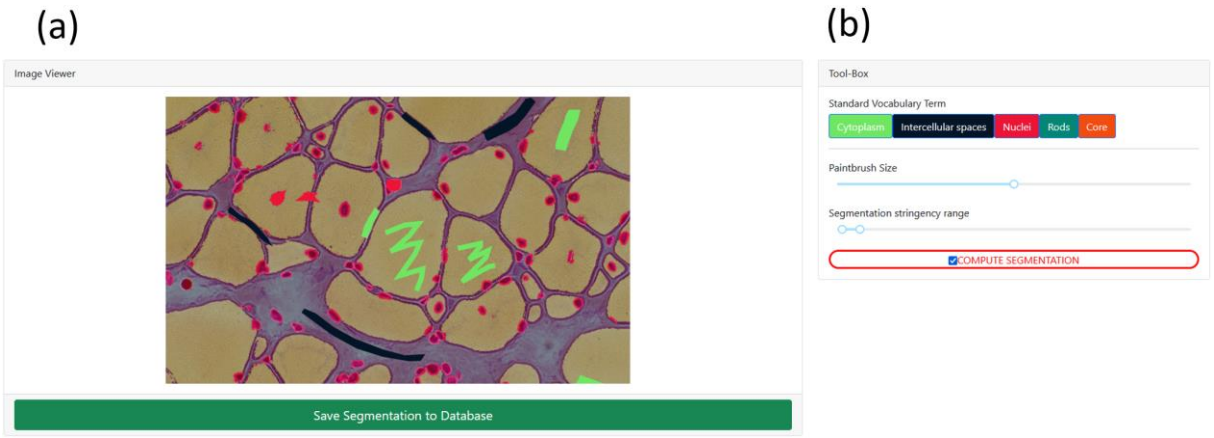


Figure 6: Screenshot of the image annotation module. (a) Image viewer used to navigate, zoom and annotate the histology image. **(b)** Menu interface to select the annotation label, brush width and segmentation parameters.

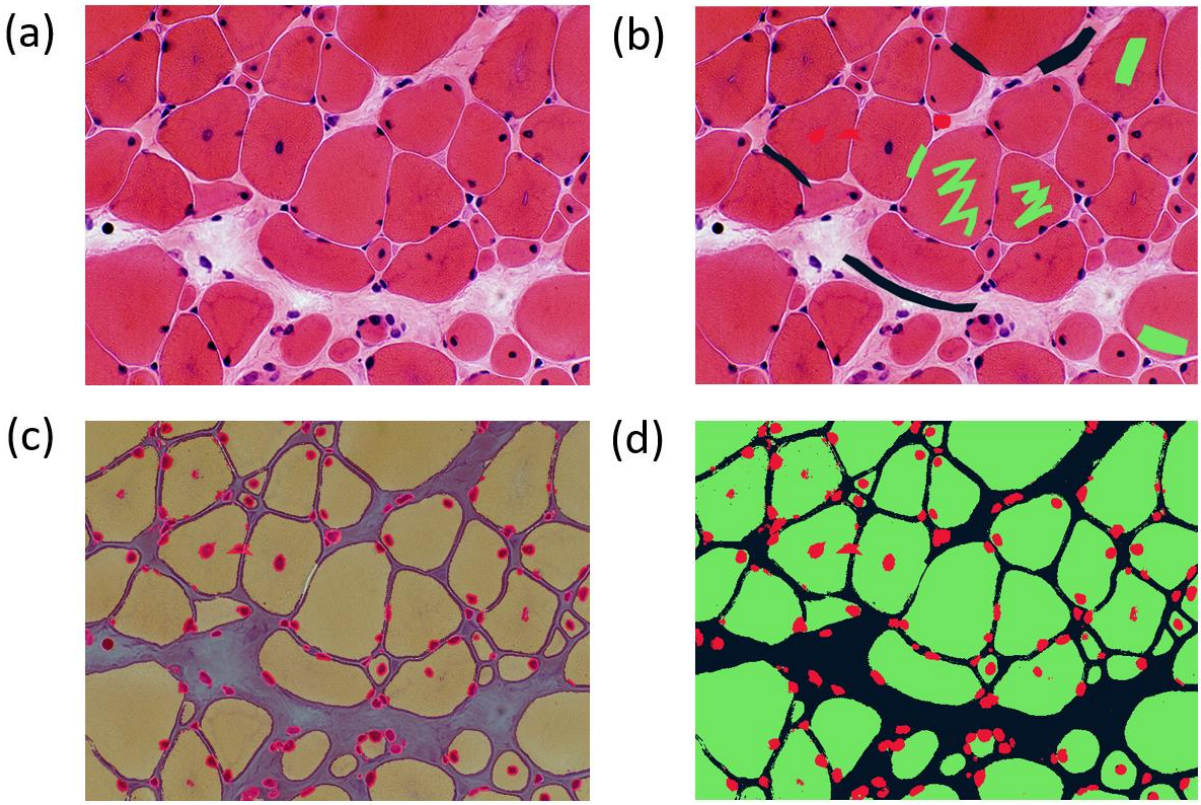


Figure 7: Image segmentation process in the image segmentation module. (a) Raw image input before annotation. **(b)** Image with limited manual annotation of cytoplasm (green), cell nucleus (red) and intercellular space (black). **(c)** Blended image of the raw image and segmented image after automated segmentation with a random-forest classifier. **(d)** Segmented image mask alone.

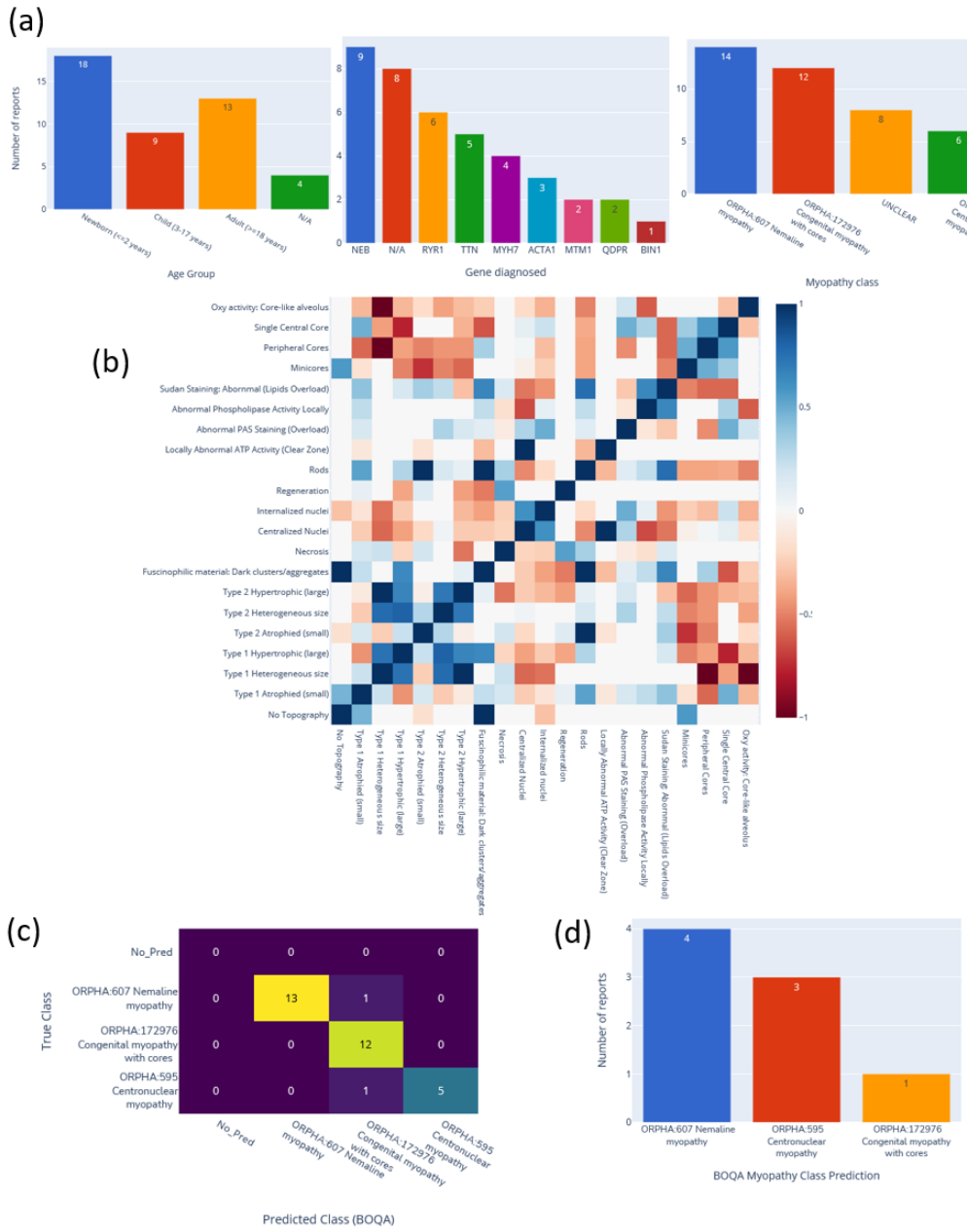


Figure 8: Automatic visualization of 40 generated congenital myopathy reports. (a) Histogram of the number of reports by age group, by diagnosed gene (top 9) or by congenital myopathy class. **(b)** Correlation matrix of standard vocabulary terms after annotation for all reports. **(c)** Confusion matrix of BOQA algorithm performance for suggestion of the three main congenital myopathy classes (NM, COM, CNM, n=32). Colors indicate the number of reports for each cell of the matrix, the lighter the color the more reports. **(d)** Histogram of the reclassification by BOQA of reports without a final diagnostic (n=8).

Table 1: Functionalities of IMPatienT compared to common state-of-the-art tools.

Group	Functionalities	IMPatienT	Phenotips	PhenoStore	SAMS	Protégé	Doc2HPO	Cytomine	Ilastik	INTEGRO
General Application Characteristics	Web application	X	X	X	X		X	X		
	Patient database	X	X	X	X					
	Free to use and open-source	X			X	X	X	X	X	X
	Support multimodal data	X								
	Support for patient pedigree data		X	X						
Standard Vocabulary	Vocabulary Builder	X				X		X		
	Advanced vocabulary terms definition	X				X				
	Full-featured ontology builder					X				
Report digitization	Integrates reference ontologies (HPO, Orphanet)	X	X		X		X			X
	Form for patient medical report digitization	X		X	X					
	Text recognition with OCR	X								
	Text processing with NLP	X					X			
	Export data to Phenopacket format			X	X					
Image annotation	Image annotation and segmentation with AI	X						X	X	
	Support for DICOM and whole slide images							X		
Patient data exploitation	Automatic visualization dashboard	X		X						
	Diagnosis prediction system	X	X							
	Data mining of information for specific diagnosis	X								X

Supplementary Materials

(a)

HISTOLOGY REPORT OF MUSCLE BIOPSY Patient Name: DOE John Patient Age: 7 years old, born on 20/11/2015 Biopsy date:
20/11/2021 Biopsy Number: 777-07
Sent on: 21/11/2021 Muscle of biopsy:
Quadriceps Cryostat sections of the fragment frozen at -160°C Hematein-eosin and Gomori trichrome.
- Two samples were analyzed.
The biopsy looks abnormal.
Muscle fibers have unequal size.
There are two populations of muscle fiber with different sizes, one is normal, the other is atrophied.
- Most muscle nuclei are in normal situations but some atrophied fibers have internalized nuclei and a few centralized nuclei.
- In Gomori trichrome, most atrophied fibers have a dark coloration with small structural reorganization.
Some fibers have a purplish shade.
- Animportant number of fibers contains fuscinophilic inclusions like rod and dark clusters, mostly in atrophied fibers.
- There seem to be a few necklace fibers.
- There is no sign of necrosis but some fibers are in regeneration.
- No increase of the interstitial connective tissue.
ATP Staining.
There is an abnormal fiber differentiation and no fiber bundling.
Oxidative Staining.
There are no fiber cores, lobulation or dark circles.
CONCLUSIONS:
The pathologic profile of this patient is similar to congenital myopathy with a nemaline subtype with strong fiber type disproportion and smaller fiber type I.
Dr. Jane Doe

(b)

Vocabulary ID	Vocabulary Term	Position In Text	Raw Text	Similarity Score
MHO:000013	unequal size	363	unequal size	100
MHO:000174	normal situation	509	normal situations	97
MHO:000027	centralized nuclei	558	internalized nuclei	86
MHO:000028	internalized nuclei	558	internalized nuclei	100
MHO:000107	structural disorganisation	668	structural reorganization	86
MHO:000048	rods	821	rod	86
MHO:000076	atpase staining	1039	atp staining	89
MHO:000001	tg staining	1039	atp staining	87
MHO:000004	fiber type 1	1353	fiber type I	92
MHO:000008	fiber type 2	1353	fiber type I	92
MHO:000074	necrosis	938	necrosis	100
MHO:000063	interstitial connective tissue increase	1006	interstitial connective tissue	87
MHO:000078	abnormal differentiation	1065	abnormal fiber differentiation	89
MHO:000123	core	1158	cores	89
MHO:000100	dark circles	1179	dark circles	100

Supplementary Figure S1: Qualitative assessment of the performances of the NLP method for matching text to the standard vocabulary. **(a)** Raw muscle histology report text with detected keywords highlighted in green and red. A red highlight indicates that the keyword is in a negated sentence. **(b)** Table of some highlighted keywords and the details of the match (matching vocabulary ID and terms, position in the raw text, matching n-gram [raw text] and the similarity score of the comparison). Green and red colors correspond to keywords detected as present or present in negated sentence respectively.

480 • **Table S1** – TableS1_frequencies_per_gene.csv - **Table of frequencies of standard vocabulary per**
481 **genes.** The CSV file contains all frequencies of standard vocabulary terms for each gene with the total
482 number of reports per gene and the number of occurrences of each term if not 0.

483 • **Table S2** – TableS2_frequencies_per_diag.csv - **Table of frequencies of standard vocabulary per**
484 **diagnosis.** The CSV file contains all frequencies of standard vocabulary terms for each diagnosis with
485 the total number of reports per diagnosis and the number of occurrences of each term if not 0.

486

5.2 Données sensibles et déploiement de la plateforme

Afin de traiter les données de patients atteints de MC (qui doivent rester privées), mais aussi de pouvoir faire une démonstration technique de l'outil, nous avons mis en ligne deux instances de la plateforme. Une première instance publique, à l'adresse https://impatient.lbgi.fr, contient une quarantaine de rapports de comptes rendus fictifs générés aléatoirement. Cette instance est accessible à tout le monde et est remise à zéro chaque jour. Une seconde instance privée, à l'adresse https://myoxia.lbgi.fr, contient 89 rapports de comptes rendus de patients provenant de l'Institut de Myologie de Paris qui ont été numérisés. Cette instance n'est accessible que par mot de passe et son contenu est sauvegardé de manière régulière. Le code source d'**IMPatienT** est *open-source* et disponible dans un répertoire GitHub à l'adresse : https://github.com/lambda-science/IMPatienT.

5.3 Limitations et perspectives de développement

IMPatienT est une application web permettant d'annoter et d'explorer les données biomédicales issues de la biopsie musculaire de patients atteints de MC. Cette application web a été développée pour pallier au manque d'outils permettant d'annoter et d'extraire de l'information de comptes rendus médicaux en texte libre grâce à une approche basée sur les ontologies. En plus d'intégrer les ontologies médicales déjà existantes (ORDO, HPO, HUGO, HGVS), **IMPatienT** permet de créer facilement un vocabulaire standard similaire à une ontologie pour les domaines où il n'existe pas encore d'ontologie de référence. Dans notre cas, nous l'avons utilisé pour créer notre propre vocabulaire standard des observations histopathologiques réalisées dans les biopsies musculaires. Nous avons passé en revue l'ensemble des 89 comptes rendus de biopsies musculaires et nous avons extrait et hiérarchisé par colorations les termes uniques trouvés dans ces rapports. Ce travail réprésente un total de 175 termes extraits composant notre vocabulaire standard.

Cependant, l'approche développée conçue pour l'annotation est semi-automatique et requiert donc toujours un travail manuel et humain de correction et de validation des annotations. Cette limitation empêche donc le passage à l'échelle et le traitement d'une masse importante de comptes rendus textuels ou d'images. De plus, l'approche utilisée est dépendante de la définition d'un vocabulaire standard exhaustif. Si un terme nouveau est présent dans un compte rendu et est absent du vocabulaire au moment de l'annotation, il faut le rajouter au préalable. De même, si on opère des modifications importantes du vocabulaire standard, il est peut-être nécessaire de devoir passer en revue l'ensemble des comptes rendus déjà numérisés pour vérifier si de nouveaux termes sont à associer aux comptes rendus ou si d'anciennes annotations sont à supprimer.

En termes de développement futur, il est nécessaire d'intégrer à **IMPatienT** de nouveaux outils permettant l'automatisation de l'analyse des comptes rendus, par exemple un outil permettant de préremplir les formulaires avec les informations générales détectées. Ainsi, grâce aux récentes avancées dans le traitement de texte libre, notamment grâce aux LLMs, nous avons développé de nouvelle méthode qui permettent d'automatiser ce processus d'annotation. Ces nouvelles méthodes sont décrites dans le chapitre 7 : "NLMyo : Traitement de rapports textuels par LLMs". Il serait intéressant aussi de proposer des alternatives plus automatiques au système d'annotation avec le vocabulaire standard.

De plus au niveau de l'analyse d'images, il est intéressant de proposer un outil capable de quantifier grâce à l'IA des marqueurs pathologiques dans les images de biopsie, car cette information est manquante dans les comptes rendus de biopsie, les observations ne sont que

qualitatives et non quantitatives. Pour cela, nous avons développé un outil présenté dans le chapitre 8 "Vers une génération de rapports automatiques à partir d'imagerie avec MyoQuant".

Grâce à l'intégration de ces outils, l'application web **IMPatienT** deviendrait le socle de l'intégration de plusieurs outils d'IA pour l'analyse de données multimodales. **IMPatienT** serait alors le point d'entrée mettant à disposition des outils pour créer une base de données multimodales de patients et fournir les outils adaptés à leur analyse et exploration.

Chapitre 6

Analyse de la base de données d'IMPatienT par IA Explicable (xAI)

La création d'un vocabulaire standard décrivant les observations (termes) faites lors de la biopsie musculaire a permis grâce à **IMPatienT** la numérisation de 89 rapports de biopsie de patients atteints de MC. En utilisant **IMPatienT** comme outil pour structurer les comptes rendus de biopsie en texte libre sous forme de données annotées, nous pouvons alors utiliser des méthodes de ML traditionnel pour entraîner des modèles prédictifs. Bien que les méthodes de ML traditionnelles soient restreintes à l'analyse de données bien structurée et annotées, elles présentent l'avantage de pouvoir être utilisable même avec une quantité de données faible (ce qui est le cas dans les maladies rares). De plus, ces méthodes présentent une meilleure explicabilité que les méthodes à base de réseau de neurones profonds. L'explicabilité d'un modèle prédictif destiné à l'aide au diagnostic est cruciale pour permettre d'avoir confiance et d'évaluer une suggestion de diagnostic automatique en médecine. Ainsi dans ce chapitre, nous avons voulu comparer les performances de plusieurs modèles de ML avec une exploration plus spécifique d'un modèle de LCS considérée comme référence parmi les algorithmes explicatifs en entraînant ces modèles sur la base de données d'**IMPatienT**.

6.1 Contenu de la base de données

Les 89 rapports d'histologie ont été numérisés en annotant *via* **IMPatienT** chacun des termes du vocabulaire entre : absence du terme (0), présence (suivant 4 gradations de quantité de 0,25 à 1) ou absence d'information (N/A). Cette procédure a permis d'obtenir une matrice numérique de taille (89x175) où les 175 colonnes représentent les 175 termes du vocabulaire standard. De plus, pour chacun des 89 rapports, un certain nombre de métadonnées ont été enregistrées telles que : l'âge, le numéro de biopsie, le muscle, le diagnostic final, l'identifiant de la biopsie, l'identifiant unique du patient.

6.1.1 Analyse statistique exploratoire

Dans un premier temps, j'ai réalisé une analyse statistique exploratoire de ce jeu de données (figure 6.1 et figure 6.2). Ainsi, grâce à **IMPatienT**, nous avons généré plusieurs visualisations dont des histogrammes (fig. 6.1, des matrices de corrélation (fig. 6.2), des tableaux de fréquences

et des matrices de confusion pour évaluer la méthode *Bayesian Ontology Query Algorithm* (BOQA). On observe dans notre jeu de données une majorité de patients adultes (49 sur 89). Au niveau des gènes, les gènes TTN, NEB, RYR1 et MYH7 sont les plus souvent responsables de la maladie (40 sur 89). Au niveau des maladies, les COM et les NM sont les plus communes (28 et 25). Ensuite, on remarque un grand groupe de 23 patients sans diagnostic clairement établi et un groupe plus restreint de 10 patients à CNM. Le tableau des termes les plus fréquents par diagnostic (accessible en ligne : https://lbgi.fr/~meyer/stat_per_diag.xlsx et dans l'archive Zenodo) fait ressortir des termes attendus tels qu'une forte fréquence de bâtonnets sombres dans les NM, de noyaux centralisés dans les CNM et de prédominance de fibre de type 1, de taille inégale des fibres et de cores (cores centraux uniques et cores périphériques) dans les COM.

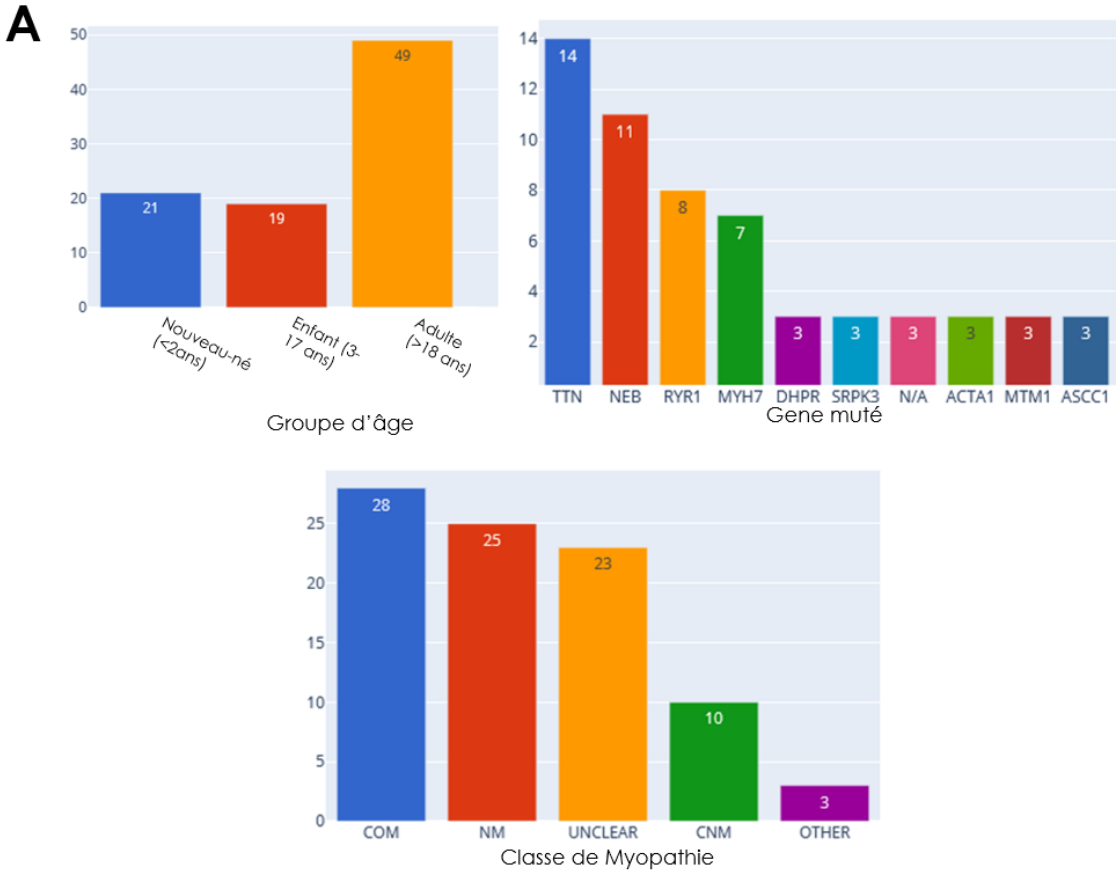


FIGURE 6.1 – **Analyse statistique exploratoire de la base de données IMPatienT** : Histogrammes de répartition des patients en fonction de leur groupe d'âge, du gène muté responsable de la maladie et de leur classe de myopathie (némaline, core, centronucléaire, autre (OTHER) et pas de diagnostic établi (UNCLEAR).

6.2 Pipeline de Machine-Learning Streamline

Nous avons voulu évaluer s'il était possible de prédire le diagnostic des patients *via* des techniques de classification par algorithmes de ML traditionnels. Pour cela, nous avons utilisé et

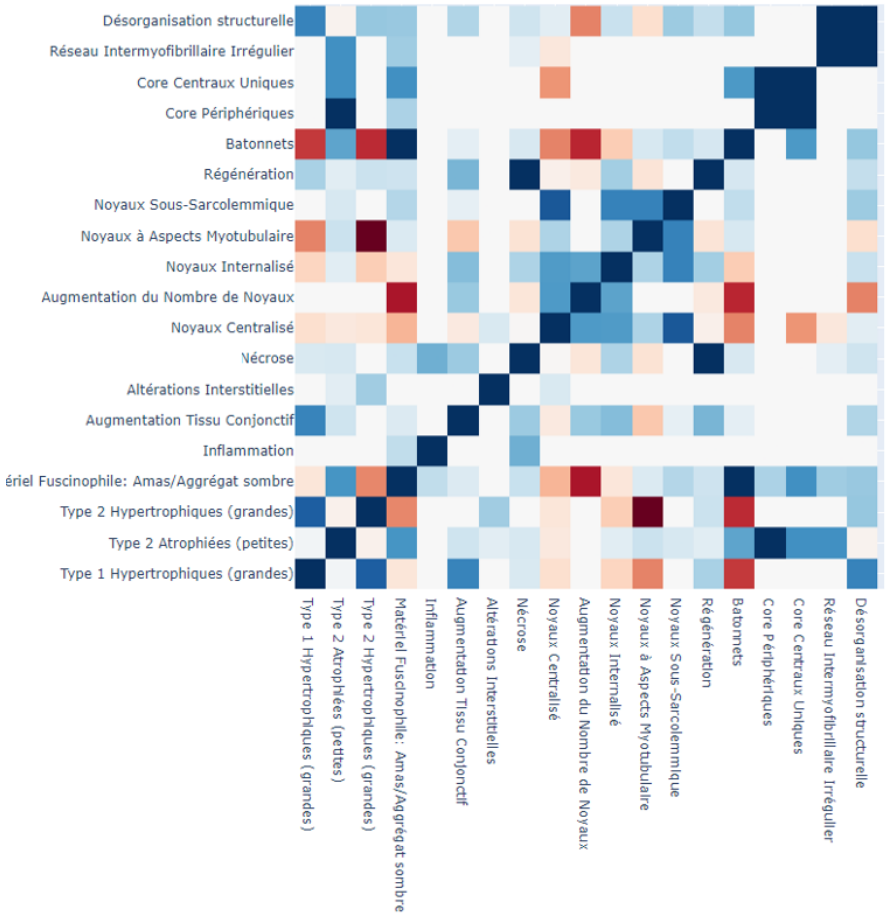


FIGURE 6.2 – **Analyse statistique exploratoire de la base de données IMPatienT** : Matrice de corrélation partielle des termes du vocabulaire standard. La couleur rouge indique une corrélation négative (-1), la couleur bleue une corrélation positive (+1). Deux termes avec une corrélation positive sont deux termes qui ont une forte co-occurrence chez les patients.

modifié le pipeline Streamline (R. Urbanowicz et al., 2023), développé par *Ryan J. Urbanowicz*, créateur aussi de l'algorithme de classification explicable (de type LCS) nommé ExSTraCS 2.0 (R. J. Urbanowicz et Moore, 2015). Le pipeline Streamline permet d'entraîner, d'optimiser et de comparer 11 algorithmes traditionnels de ML sur un même jeu de données. Ce pipeline est développé pour la classification binaire, j'ai donc procédé à des modifications pour le rendre compatible avec des tâches de classification multi-classes (prédiction de diagnostic entre les NM, COM et les CNM).

6.3 Résultats d'analyse

L'utilisation de Streamline nous a permis de comparer 11 algorithmes de ML pour la classification de données biomédicales (tableau 6.1). Les 11 algorithmes ont montré une exactitude de classification (métrique qui considère le pourcentage de classification correcte au global) se situant entre 0,71 et 0,86. Si l'on regarde les performances pour le coefficient de corrélation de Matthew (la seule métrique de performance prenant en compte l'ensemble de la matrice de confusion) l'écart entre les algorithmes est encore plus faible. Dix des onze algorithmes ont un score MCC compris entre 0,72 et 0,79. Au niveau de la courbe ROC (comparaison du taux de vrais positifs par rapport aux faux positifs pour différents seuils), les algorithmes semblent similaires avec une aire sous la courbe (*area under the curve*, AUC) légèrement en deçà des autres algorithmes pour ExSTraCS et l'arbre de décision (aux alentours de 0.85 contre 0.90 et 0.95 pour les neuf autres, 6.3).

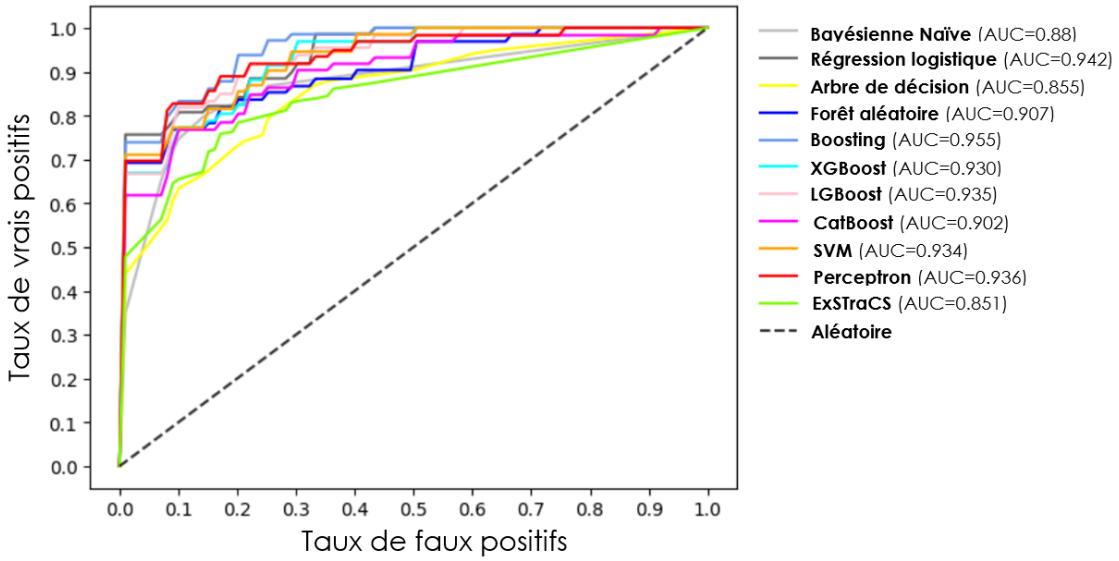


Figure 6.3 – **Comparaison des courbes ROC des 11 algorithmes comparés**. La valeur AUC représente l'aire sous la courbe, plus elle est proche de 1 plus le modèle est performant.

Les algorithmes *Support Vector Machine* (SVM) et *Boosting* se sont révélés être les meilleurs algorithmes en termes d'exactitude équilibrée et de score MCC, cependant ces deux modèles ne sont pas des modèles transparents et nécessitent de méthodes *post-hoc* pour interpréter les prédictions (explicabilité). L'algorithme ExSTraCS, qui lui est un modèle totalement transparent et explicable, a obtenu des performances 22% inférieures au meilleur modèle, avec un score MCC

Algorithme	Exactitude Équilibrée	Exactitude	Matthew Corr. Coeff.	Spécificité
Bayesienne Naive	0.67	0.82	0.74	0.91
Regession Logistique	0.86	0.82	0.72	0.91
Arbre de décision	0.72	0.72	0.74	0.86
Forêt Aléatoire	0.81	0.82	0.74	0.91
Boosting	**0.89**	0.85	0.78	0.92
XGBoost	0.75	0.82	0.74	0.91
LGBoost	0.67	0.83	0.76	0.92
CatBoost	0.67	0.82	0.73	0.91
SVM	**0.89**	0.85	**0.79**	0.92
Perceptron	0.67	**0.86**	0.78	**0.93**
ExSTraCS	0.61	0.71	0.61	0.86

TABLEAU 6.1 – **Comparaison des performances des modèles (moyenne sur 10 *folds* de cross-validation)**. Les méthodes SVM et Boosting ont obtenu les meilleures performances à travers la majorité des métriques. La méthode LCS ExSTraCS a obtenu des performances 22% inférieurs aux meilleurs modèles.

de 0,61 et une exactitude de 0,71, ce qui est constamment en dessous des autres algorithmes. Cette performance plus faible peut s'expliquer en partie par les contraintes lors de son entraînement lié à l'explicabilité. En effet, nous avons volontairement restreint le nombre de règles que l'algorithme peut générer pour le rendre plus facilement interprétable et pour pouvoir en extraire les principales règles de classification entre les sous-types de myopathies (voir section 6.4). Il est intéressant de noter que l'algorithme de classification d'arbre de décision, qui est à la fois simple et par nature explicable, a obtenu des performances dans la moyenne des autres algorithmes, tout en ayant un temps d'entraînement faible de 17 secondes en raison de sa simplicité (tableau 6.2). À titre de comparaison, le LCS ExSTraCS a nécessité un temps d'entraînement de 5min39, et ceci sans phase d'optimisation des hyperparamètres.

Algorithme	Temps d'optimisation et d'entraînement (sec)
Bayesienne Naive	11
Régression Logistique	30
Arbre de décision	17
Forêt Aléatoire	2072
Boosting	1887
XGBoost	733
LGBoost	168
CatBoost	4578
SVM	22
Perceptron	389
ExSTraCS	339

TABLEAU 6.2 – **Temps d'optimisation des hyperparamètres et d'entrainement des algorithmes de ML**. Certains modèles simples comme l'arbre de décision, la régression logistique et le modèle SVM ont un temps d'entrainement et d'optimisation inférieur à 1 minute. Les méthodes complexes comme le Boosting, la forêt aléatoire et le LCS ont des temps d'entrainement et d'optimisation supérieurs à 30 minutes.

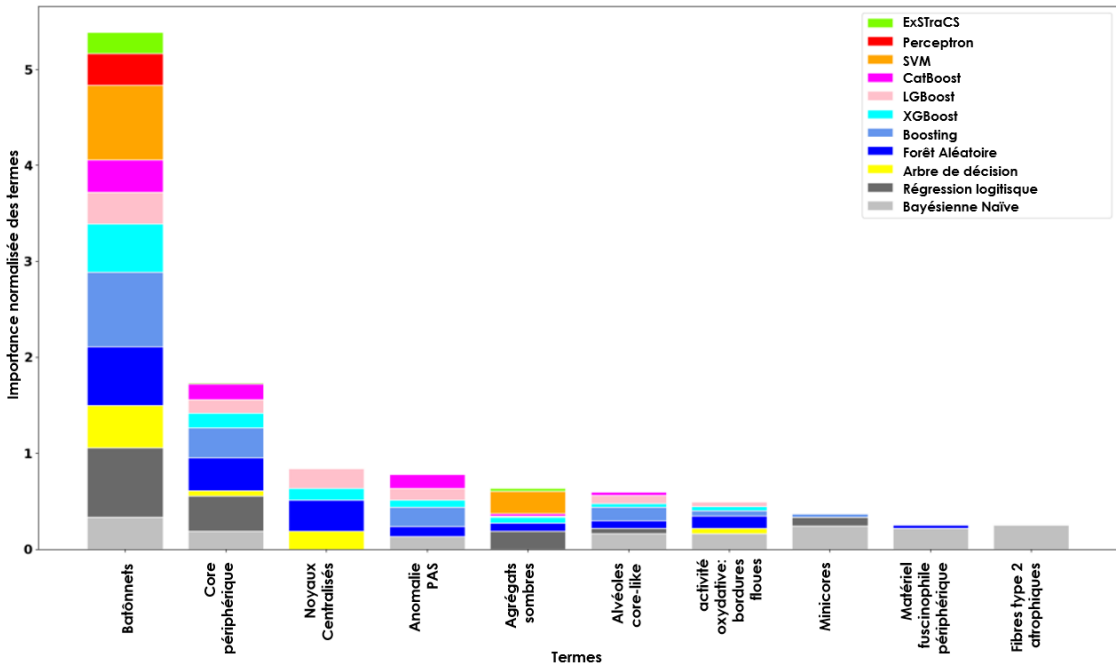


FIGURE 6.4 – **Histogramme des 10 termes les plus déterminants pour la classification pour chaque algorithme considéré**. Une très grande importance est accordée par tous les algorithmes à la présence de bâtonnets pour faire la différence entre les sous-types de myopathies.

L'histogramme montrant les 10 termes les plus importants pour la classification de toutes les classes de MC dans l'ensemble des algorithmes (fig. 6.4) mets en évidence des termes attendus comme la présence de bâtonnets, de cores ou de noyaux centralisés pour différencier NM, COM et CNM avec un poids très important pour la présence de bâtonnets. Mais on observe aussi la présence de critères moins attendus, comme la présence d'anomalie au marquage PAS ou l'atrophie des fibres de type 2. De plus on observe aussi que tout les algorithmes n'accordent pas la même importance aux mêmes termes. Par exemple, la présence de fibres de type 2 atrophiques ne semblent importantes que pour la méthodes bayésienne naïve.

La faible taille de notre jeu de données (89 patients), sa grande dimensionalité et hétérogéinité (175 termes) couplé aux faibles différences en termes de performances entre les algorithmes semble restreignantes concernant la comparaison des algorithmes. Sur notre jeu de données, il semblerait que l'utilisation d'algorithmes simples comme l'arbre de décision soit à privilégier pour sa triple efficacité en termes de performances, explicabilité et coût en puissance de calcul. Il est possible que l'utilisation d'un ensemble de données plus important, c'est-à-dire contenant plus de comptes rendus de biopsies, et plus homogènes permettrait une meilleure évaluation des performances de ces algorithmes et, par conséquent, une meilleure compréhension de leurs avantages et inconvénients respectifs dans le contexte de la classification de données biomédicales de manière explicable.

6.4 Méthode de visualisation des règles de LCS

Bien que notre algorithme de LCS présente des performances inférieurs aux autres algorithmes en terme de classification, ils ont l'avantage de produire produisent une liste de règles pour la classification, ce qui rend le processus de prédiction totalement transparent. Cette liste de règles explicites, peut permettre d'utiliser cet algorithme comme extracteur de connaissances à partir de nos données, en découvrant des règles pertintentes de classification qui n'étaient pas connu à priori.

Cependant, une simple liste de règles sous la forme d'un tableau ne permet pas d'extraire des connaissances facilement et rapidement de ces modèles. Nous avons voulu explorer différentes approches pour représenter graphiquement ces règles, afin d'en extraire des connaissances de manière visuelle. Le code écrit pour générer ces visualisations est open source et est disponible dans un répertoire GitHub à l'adresse : https://github.com/lambda-science/PredEx.

6.4.1 Principe général

Nous avons développé deux approches pour la visualisation des règles produites par l'entrainement de ExSTraCS. Au total, ce sont 42 règles accessibles en ligne à l'adresse https://lbgi.fr/~meyer/lcs_rules.xlsx ou dans l'archive Zenodo associée qui ont été générées.

La première approche consiste à visualiser les interactions entre les termes dans les rapports, c'est-à-dire leur co-occurrence dans les règles générées. Lorsque deux termes apparaissent dans une même règle définie par un LCS, un lien est tracé entre eux, produisant un diagramme de cordes. Un lien épais entre deux termes indique une co-occurrence importante dans la liste des règles, suggérant une relation étroite entre les termes du vocabulaire standard des biopsies musculaire. La seconde approche consiste à visualiser les liens entre les termes et les diagnostics. Sur un graphe, chaque nœud circulaire représente un terme, tandis qu'un nœud triangulaire représente un diagnostic. Pour chaque règle, un lien est tracé entre les termes et le diagnostic correspondant. Plus le terme apparait dans des règles liées à un diagnostic, plus le lien sera épais, indiquant une relation forte entre l'observation et la classe de MC. Les liens sont colorés en rouge ou vert en fonction de l'absence ou la présence du terme menant à la classe de MC, respectivement. Par exemple, si toutes les règles liant le terme "core" aux NM stipulent que les cores doivent être absents, alors le lien sera rouge. Si l'observation est liée au diagnostic à la fois en état d'absence et de présence dans des règles différentes, le lien est coloré en jaune, reflétant la complexité de la relation entre l'observation et le diagnostic.

6.4.2 Résultats

La première approche a permis de générer le diagramme de cordes présenté dans la figure 6.5. Cette figure est disponible de manière interactive à l'adresse https://lbgi.fr/~meyer/chord.html et dans l'archive Zenodo. Sur cette représentation, nous avons filtré les liens pour ne conserver que les liens ayant une valeur supérieure ou égale à 3 (co-occurrence des deux termes dans au moins 3 règles). On observe sur cette figure des interactions complexes et multiples entre les termes. Par exemple, les règles impliquant le terme"Type 1 atrophiées" incluent toujours le terme "sans topographie". De même pour le terme "Type 2 hypertrophiques" et le terme "Activité oxydative : alvéoles core-like", qui apparaissent conjointement dans 4 règles différentes.

La seconde approche a permis de générer le réseau représenté dans la figure 6.6. Cette figure est disponible de manière interactive à l'adresse https://lbgi.fr/~meyer/myomap.html et dans l'archive Zenodo. On observe sur cette représentation que chaque diagnostic est lié à des termes exclusifs tel que la présence de nécrose et d'une augmentation du collagène endosymial pour les

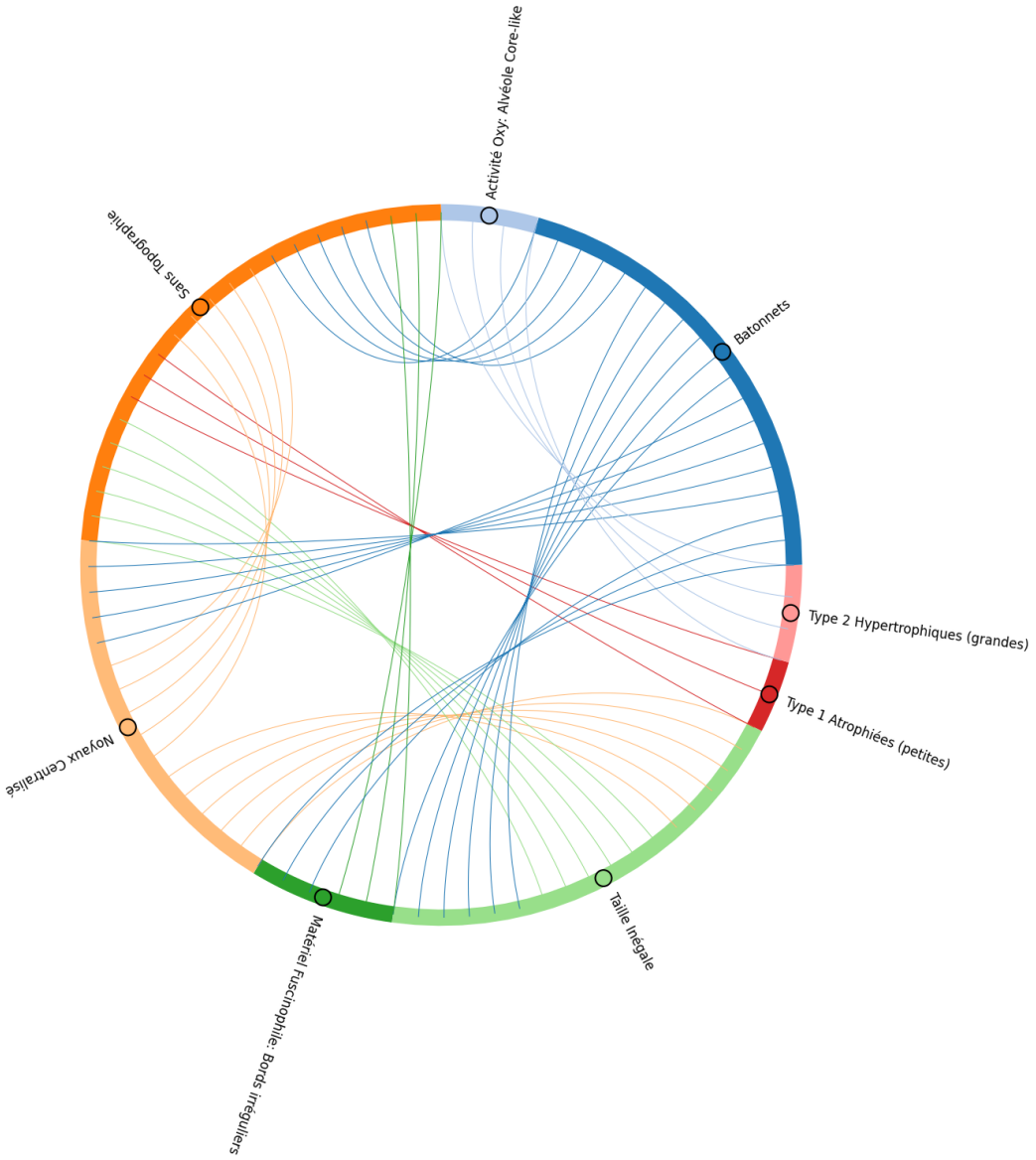


FIGURE 6.5 – **Diagramme de cordes des règles issues de l'entraînement de ExSTraCS**. Un lien entre deux termes indique que les deux termes apparaissent de manière commune dans une règle créée par le modèle. Plusieurs liens entre deux termes indiquent que ces liens apparaissent de manière commune dans plusieurs règles distinctes.

myopathies à némaline ou un trouve de l'activité phospholipases pour les myopathies à cores. On observe aussi que trois termes sont utilisés par les modèles pour discriminer entre les sous-types de myopathies en fonction de leur état. Par exemple, le terme "type 2 hypertrophiques" est lié de façon positive aux myopathies à cores et négativement aux myopathies à némaline et centronucléaires. À l'inverse, le terme bâtonnet est lié positivement aux myopathies à némaline, négativement aux myopathies à cores et de façon mixtes (jaune) aux myopathies centronucléaires. Enfin, on observe que cette visualisation permet de mettre en évidence de nouveaux termes présentés comment pertinents pour la classification des myopathies qui n'avaient pas été mis en évidence dans l'histogramme de l'importance des termes pour la classification présenté plus haut (fig. 6.4).

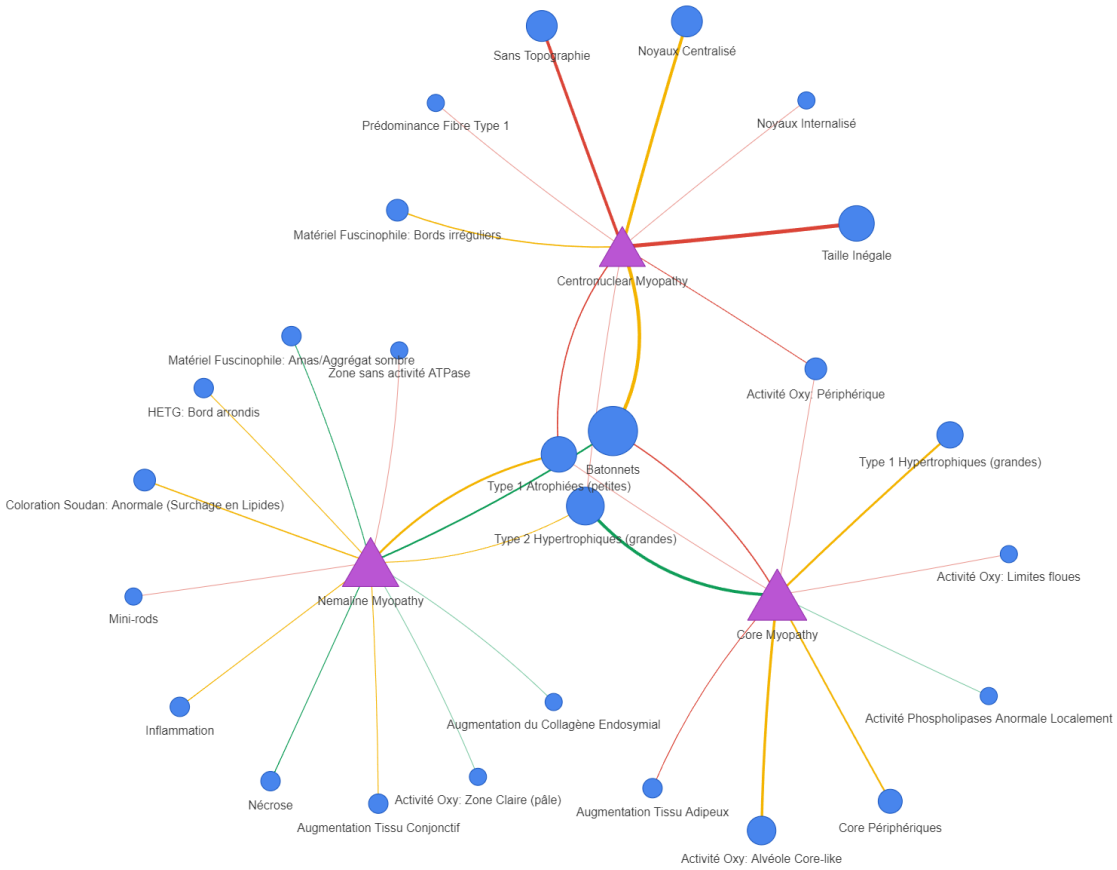


FIGURE 6.6 – **Représentation des règles issues de l'entraînement de ExSTraCS sous forme de réseau**. Les triangles violets indiquent une classe (diagnostic de myopathie). Les nœuds bleus sont des termes du vocabulaire standard. Un lien vert indique un lien positif entre la classe et le terme, c'est-à-dire que le terme n'apparait que dans des règles liant sa présence au diagnostic. Un lien rouge indique que le terme n'apparait que dans des règles liant son absence au diagnostic. Un lien jaune indique des règles pouvant utiliser la présence ou l'absence du terme.

6.5 Perspectives de développement

En termes de développement futur, il est nécessaire d'agrandir le jeu de données utilisé pour comparer les algorithmes de classification. En effet, notre jeu de données de 89 rapports semble trop hétérogène et de trop petite taille pour déceler une différence significative de performances entre les algorithmes. De plus, il est nécessaire d'améliorer les approches de visualisation des règles de LCS, car ces approches semblent être intéressantes pour visualiser graphiquement et rapidement le fonctionnement interne de notre système de classification explicable et pour identifier de nouveaux critères pertinents pour différencier les sous-types de myopathies congénitales.

Enfin, il pourrait être intéressant d'intégrer ce *pipeline* d'entraînement et ces visualisations à **IMPatienT** pour entraîner automatiquement un modèle de classification explicable à base de règles et performant lors de l'entrée de nouveaux patients. De plus, l'intégration de ce modèle dans le formulaire de numérisation de **IMPatienT** pourrait permettre de réaliser de l'aide au diagnostic en suggérant automatiquement un diagnostic et en mettant en évidence les règles à l'origine de cette prédiction.

Chapitre 7

NLMyo : Traitement de rapports textuels par LLMs

Dans les deux précédents chapitres, nous avons présenté **IMPatienT** un outil d'annotation et d'exploration de compte rendu de biopsie en texte libre. **IMPatienT** utilise un système à base d'ontologie et de vocabulaire standard pour détecter et annoter la présence ou l'absence d'éléments pathologiques dans les biopsies musculaires. Cependant, ce système présente certaines limites. Tout d'abord, il requiert de créer un vocabulaire standard exhaustif pour décrire les observations dans les biopsies musculaires, ce qui est un travail manuel important. De plus, le système d'annotation semi-automatique utilise un système à base de règles et de correspondance exacte des mots aux ontologies existantes. Cette correspondance exacte réduit la flexibilité du système et sa sensibilité de détection, il faut alors réaliser un travail d'annotation manuel important pour numériser les comptes rendus de biopsie.

En fin 2022 et début 2023, les récentes avancées dans le domaine du NLP ont permis de révolutionner la manière de traiter et d'exploiter les données sous forme de texte libre. La mise à dispositions de modèles linguistiques de grande taille (LLMs, 2.3.5) performants, accessibles et capables de suivre des instructions, ouvre la porte à la création d'outils plus performants et flexibles pour le traitement de ces comptes rendus. Ces systèmes basés sur une approche sémantique et multilingue éliminent la nécessité de définir un vocabulaire standard. Ainsi nous avons développé **NLMyo** (fig 7.1), une boite à outils basée sur les LLMs mettant à disposition quatre outils généralistes pour le traitement de comptes rendus médicaux : outil d'anonymisation, d'extraction d'information, de classification automatique et de création de moteurs de recherche. L'ensemble de ces outils et des modèles utilisés est représenté dans la figure 7.2.

7.1 *Anonymizer* : un outil d'anonymisation

Le premier outil de **NLMyo** est *Anonymizer*, un outil permettant de supprimer automatiquement les informations identifiantes des comptes rendus médicaux. Dans les comptes rendus de biopsie de l'Institut de Myologie de Paris que nous traitons, deux données identifiantes et personnelles sont présentes et doivent être retirées : le nom du patient (et du personnel médical) ainsi que la date de naissance du patient. Non seulement ces informations ne sont pas utiles pour les analyses subséquentes, mais de plus, par respect de la vie privée et des recommandations Règlement général sur la protection des données (RGPD), ces informations ne doivent être

FIGURE 7.1 – **Logo de NLMyo**. NLMyo est une boite à outils permettant d'utiliser les LLMs pour exploiter automatiquement les comptes rendus de biopsie musculaire. NLMyo est open-source et disponible en version de démonstration en ligne. https://github.com/lambda-science/NLMyo

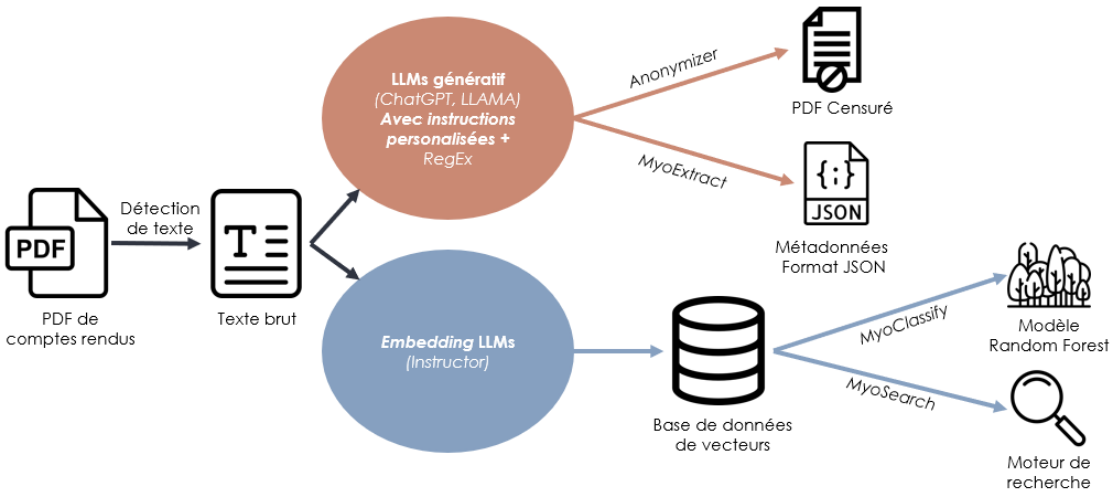


FIGURE 7.2 – **Structure de NLMyo**. NLMyo utilise deux types de LLMs pour traiter les comptes rendus de biopsie : les modèles génératifs et les modèles d'*embedding*. L'utilisation de ces deux types de modèles permet la mise à disposition de quatre outils : Anonymizer, MyoExtract, MyoClassify et MyoSearch.

accessibles qu'aux professionnels en charge du patient. L'anonymisation des rapports est donc une étape essentielle avant le transfert des données, leur numérisation et leur analyse.

Afin de traiter un grand volume de rapports et d'éliminer le travail manuel nécessaire, nous avons tenté d'automatiser la tâche d'anonymisation *via* deux approches : une approche traditionnelle par Expressions régulières (RegEx) et une approche novatrice par LLMs.

7.1.1 Anonymisation par RegEx

En première intention, nous avons développé une méthode basée sur les RegEx pour leur simplicité de mise en place et leur rapidité d'exécution (coûts en puissance de calcul faible). Les RegEx sont des séquences de caractères capables de trouver des motifs dans un texte, par exemple toutes les lignes commençant par "Nom : ". Le tableau 7.1 liste les RegEx utilisées pour capturer les informations de noms et de dates. Les comptes rendus de biopsie sont semi-structurés. Pour la plupart, le nom du patient est facilement identifiable, car il est précédé par le préfixe : "Nom : ". Ceci est facilement représenté par la première RegEx listée dans le tableau. Ensuite pour les autres cas de figure, comme les noms de famille sont souvent en majuscule et les prénoms commencent souvent par une majuscule, nous avons développé deux autres RegEx pour capturer les couples de mots dont un est en majuscule et le second commence par une majuscule (ou inversement, lignes 2 et 3 du tableau). Ensuite, une troisième RegEx a été ajoutée pour détecter les dates au format JJ-MM-AAAA ou JJ.MM.AAAA (ligne 3 du tableau). Finalement, une dernière RegEx est utilisée pour essayer de trouver le numéro de biopsie dans le document afin de renommer le fichier avec un nom unique et anonyme (ligne 4 du tableau).

Tableau 7.1 – **Expressions régulières pour extraire les noms et les dates**. Trois expressions régulières sont utilisées pour détecter les noms de patients, une expression pour les dates et une expression pour les numéros de biopsie.

Expression régulière	Syntaxe	Exemples d'utilisation
Nom patient	Nom.* : *([A-Za-zÀ-ÿ-]+)	Nom : Pierre Laroche
Nom patient 2	(([A-Z][a-zÀ-ÿ-]{3,}?)+ ([A-Z-]{3,}?)+)	LAROCHE Pierre
Nom patient 3	(([A-Z-]{3,}?)+ ([A-Z][a-zÀ-ÿ-]{3,}?)+)	Pierre André LAROCHE
Date	([(.]?[0-9]{1,2}[./][0-9]{1,2}[./][0-9]{1,4}[().]?)	07/04/1994, 07.05.18
N° Biopsie	([0-9]{3,8}[-/]?[0-9]{0,3})	7377-07, 1234/56

La figure 7.3 présente les résultats de la technique d'anonymisation par RegEx sur l'entête d'un rapport factice de patient, mais avec une structure similaire aux rapports de l'Institut de Myologie de Paris. Les noms et les dates ont été censurés correctement et la méthode a produit de bons résultats. Cependant, cette méthode est insuffisante, car elle est peu sensible (dates ou noms non censurés) et spécifique (informations considérées à tort comme des dates ou des noms). Par exemple tout en haut de l'entête, la date d'envoi "12 juin 2018" n'a pas été censurée. Le tableau 7.2 liste trois exemples de cas où cette méthode ne fonctionne pas et produit des erreurs. Il est possible de corriger ces erreurs en augmentant la somme de RegEx utilisées, mais augmenter le nombre de RegEx augmente aussi potentiellement le nombre de faux positifs. De plus, il n'est pas toujours possible de construire une RegEx adaptée pour extraire une information précise. Nous avons alors exploré la capacité des LLMs pour la recherche et l'extraction de ces informations de manière plus robuste et flexible que la méthode RegEx.

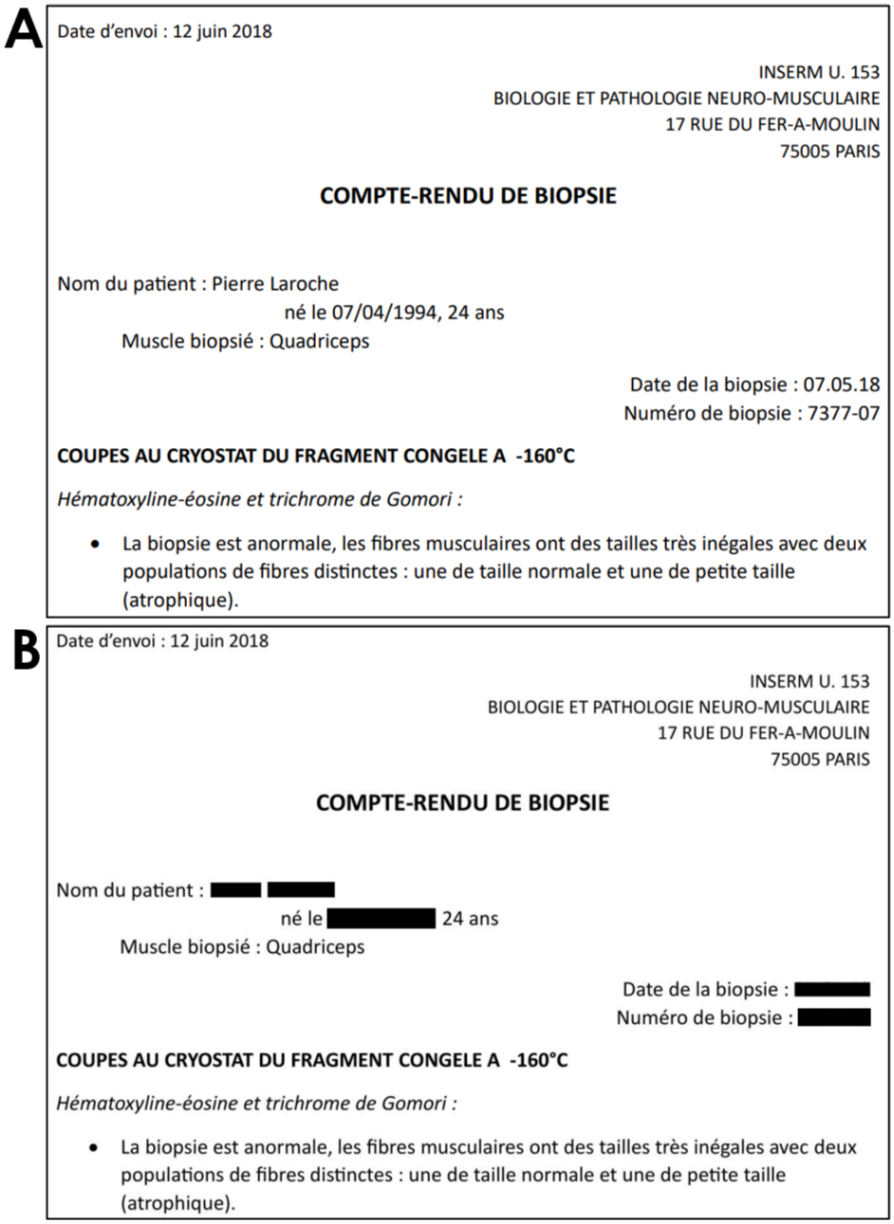


FIGURE 7.3 – **Exemple d'anonymisation d'un compte rendus fictif de biopsie avec la méthode RegEx. (A)** Le rapport à l'état brut **(B)** le rapport censuré. On observe que le nom, le numéro de biopsie et deux dates ont été censurées. Cependant la date d'envoi "12 juin 2018" n'a pas été censurée.

Tableau 7.2 – **Exemples de faux positifs ou faux négatifs de la méthode RegEx**. En fonction du cas de figure, les RegEx peuvent provoquer des faux positifs (censure d'informations excessive) ou des faux négatifs (information personnelle non censurée correctement).

Texte	RegEx déclenchée	Type	Commentaire
PAS Staining	Nom patient 3	Faux Positif	Nom de coloration dont la notation est confondue avec le motif "NOM Prénom"
Louis C. Dupont	N/A	Faux Négatif	La présence du "C." au centre ne permet pas aux RegEx de nom de se déclencher
12 mars 2001	N/A	Faux Négatif	La notation de date avec un mois en lettres ne permet pas à la RegEx de date de se déclencher
1996-04	N° Biopsie	Faux Positif	La notation de date AAAA-MM déclenche la RegEx de numéro de biopsie.

7.1.2 Anonymisation par LLMs

Les LLMs sont basés sur la compréhension du sens sémantique du texte tandis que les RegEx sont basées sur le principe de motifs de caractères et donc sur la structure du document. L'idée de l'anonymisation par LLMs est d'utiliser un modèle génératif auquel on fournit une instruction et le texte à anonymiser. L'instruction liste les informations à extraire du texte et spécifie le format de sortie. Pour intégrer ces modèles génératifs à outil d'anonymisation, nous voulons récupérer les informations extraites dans un format exploitable informatiquement, nous avons utilisé le format JSON. Le format JSON représente un dictionnaire informatique qui est utilisable par l'application pour censurer les PDF à partir des informations détectées.

7.1.3 Instruction personnalisée et *one-shot learning*

Nous avons construit une instruction personnalisée en 3 parties qui intègre une méthode de *one-shot learning*. Le *one-shot learning* est une technique d'apprentissage qui vise à généralisation en ne donnant qu'un seul exemple d'apprentissage au modèle pour la l'apprentissage d'une nouvelle tâche. Dans l'instruction les trois parties sont : (i) la description de la tâche à effectuer (ii) un exemple de réalisation de la tâche *(one-shot learning)*, (iii) le texte d'intérêt à anonymiser. Voici un exemple d'instruction que nous utilisons pour réaliser l'extraction des noms et des dates dans les rapports :

> Tu es un assistant qui extrait des informations d'un texte libre. Le format de ta réponse doit être un format JSON valide qui respecte le nom des clés fournies. Si une valeur est manquante, indique simplement N/A, n'essaie pas d'inventer. Voici la liste des informations à récupérer, les clés JSON sont indiquées entre parenthèses : nom complet (name), dates (date).
>
> ENTRÉE :
>
> Kendrick Lamar et Jane Clinton sont asymptomatiques. Date de naissance : 16 février 1991, numéro de biopsie : 666-77. Ce rapport a été expédié le 01.04.1991.
>
> SORTIE :
>
> {"name" :["Kendrick Lamar", "Jane Clinton"], "date" : ["16 février 1991", "01.04.1991"]}

ENTRÉE :
<texte à analyser>
SORTIE :

La partie précédant le mot clé "ENTRÉE" correspond à l'instruction décrivant précisément la tâche que le modèle doit réaliser (liste des informations à extraire, format de sortie et comportement attendu). Le premier couple "ENTRÉE" et "SORTIE" correspond à un exemple de réalisation de la tâche, ce qui permet de spécifier le schéma JSON attendu au modèle (*one-shot learning*). Puis le second couple "ENTRÉE", "SORTIE" correspond à l'endroit où l'on injecte notre texte d'intérêt à analyser et spécifie au modèle que l'on attend maintenant une sortie textuelle au format JSON pour l'entrée précédente.

7.1.4 Exemple et comparaison à la méthode RegEx

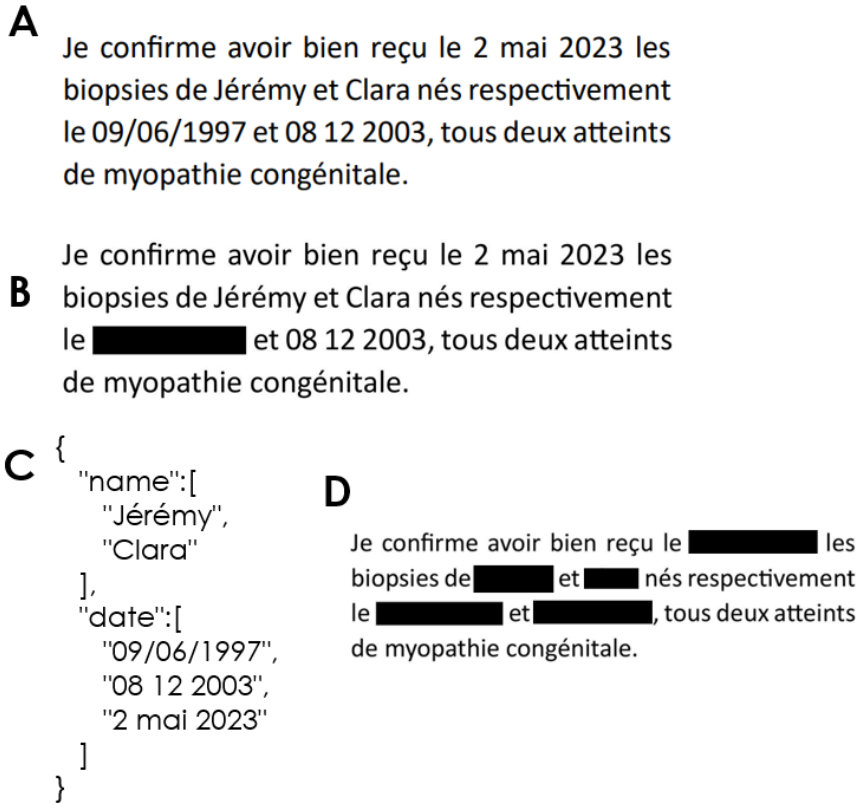


FIGURE 7.4 – **Exemple d'anonymisation d'un courrier médical factice avec la méthode par LLMs**. **(A)** Texte brut, **(B)** Anonymisation par RegEx, **(C)** JSON brut généré par le LLMs GPT-3.5-turbo (OpenAI), **(D)** Anonymisation à partir du JSON généré par LLMs.

Dans cet exemple (figure 7.4), nous avons construit un début de courrier médical factice faisant figurer des informations personnelles qui ne sont pas détectables par la méthode RegEx. Les prénoms "Jérémy" et "Clara", ainsi que les dates "2 mai 2023" et "08 12 2003" n'ont pas été détectées par la méthode RegEx et n'ont pas été censurées. On observe par contre que la méthode par LLMs (C et D) a été capable d'identifier ces informations et de les extraire du texte. Cet

exemple montre que les LLMs peuvent être la base d'un système d'anonymisation plus flexible et moins dépendant de la structure du document.

7.2 *MyoExtract* : un outil d'extraction d'information

À partir des résultats encourageants obtenus avec la technique d'anonymisation par LLMs pour l'extraction d'information, nous avons voulu étendre le champ des informations extraites de manière automatique à partir de texte libre. Nous avons utilisé la même stratégie d'extraction d'information, c'est-à-dire l'utilisation de LLMs génératifs, mais avec une instruction légèrement différente. Cette fois-ci, nous avons ajouté une liste plus importante d'informations à extraire dans le but d'en extraire les métadonnées commune à tout les comptes-rendus de biopsie. Par exemple, nous avons cherché à extraire : les noms, date de naissance, date d'envoi de la biopsie, numéro de biopsie, muscle prélevé, diagnostic final. De plus, nous avons cherché à savoir s'il était possible d'extraire les mentions d'anomalie pour certaines colorations telles que la coloration PAS, Soudan, COX, ATP et Phosphorilase. Cette extraction d'information pourrait permettre d'annoter automatiquement les rapports avec une liste d'anomalies détectées pour chaque coloration. De même que précédemment, nous avons construit une instruction personnalisée en 3 parties (description, exemple, texte à analyser).

Voici un exemple d'instruction que nous utilisons pour réaliser l'extraction des métadonnées et d'anomalies générales des colorations à partir de rapports (à noter que l'instruction présentée est en français mais fonctionne pour l'analyse de texte anglais car les modèles génératifs sont multilingues et peuvent utiliser des instructions contenant un mélange de langues) :

> Tu es un assistant qui extrait des informations d'un texte libre. Le format de ta réponse doit être un format JSON valide qui respecte le nom des clés fournies. Si une valeur est manquante, indique simplement N/A, n'essaie pas d'inventer. Formate les dates sous la forme DD-MM-YYYY et convertis les âges en années (0 si inférieur à 1 an). Voici la liste des informations à récupérer, les clés JSON sont indiquées entre parenthèses : nom complet (name), âge (age), date de naissance (birth), date de la biopsie (biodate), date d'envoi de la biopsie (sending), muscle (muscle), numéro de la biopsie (bionumber), diagnostic (diag), présence d'une anomalie dans la coloration du PAS (PAS), présence d'une anomalie dans la coloration Soudan (Soudan), présence d'une anomalie dans la coloration COX (COX), présence d'une anomalie dans la coloration ATP (ATP), présence d'une anomalie dans la coloration Phosphorylase (phospho)
>
> ENTRÉE :
>
> Kendrick Lamar et Jane Clinton ne sont pas asymptomatiques. Date de naissance : 16 février 1991, numéro de biopsie : 666-77. Anomalie forte à la coloration PAS, mais pas d'anomalie à la coloration lipide soudan. Le tableau est révélateur d'une myopathie à némaline.
>
> SORTIE :
>
> {"name" :["Kendrick Lamar", "Jane Clinton"], "age" :"N/A", "birth" : "16-02-1991", "biodate" : "N/A"", "sending" : "N/A"", "muscle" : "N/A"", "bionumber" : "666-77", "diag" : "myopathie à némaline", "PAS" : "yes", "Soudan" : "no", "COX" : "N/A", "ATP" : "N/A", "phospho" : "N/A"}
>
> ENTRÉE :
>
> <texte à analyser>
>
> SORTIE :

7.2.1 Exemple d'extraction d'information

Pour cet exemple d'utilisation, nous avons généré un rapport factice de patient avec une structure similaire aux rapports de l'Institut de Myologie de Paris qui reprend des observations typiques trouvées dans les rapports réels de biopsie. Ce rapport est disponible en figure 7.5.

Pour extraire les informations, nous avons comparé deux modèles présentés dans le chapitre 4 "Matériels et méthodes" : un modèle performant et accessible uniquement via l'API commerciale OpenAI (GPT-3.5-turbo) et un modèle auto-hébergé libre et open source *Vicuna-7B*. Les résultats de l'extraction d'information présentés dans le tableau 7.3 montrent que le modèle d'*OpenAI* GPT-3.5-turbo est capable d'extraire l'ensemble des informations demandées de manière satisfaisante sans erreurs tout en étant capable de détecter l'absence de certaines informations. Concernant le modèle *Vicuna-7B*, qui a l'avantage d'être auto-hébergé et donc d'être utilisable pour des données sensibles, les performances sont moindres. En effet, six données sur treize ont été extraites correctement (nom, date de naissance, muscle, numéro de biopsie, diagnostic, anomalie PAS). Cependant, sept autres informations demandées ont été loupées par le modèle indiquant simplement "N/A".

Il est important de noter qu'en termes de ressources de calcul et de temps d'inférence, *GPT-3.5* possède un avantage non négligeable, car ce modèle n'est accessible que par une API. Les coûts de calcul pour l'application sont donc nuls et la requête ne prend que quelques secondes à être réalisée. Pour *Vicuna-7B*, le modèle étant auto-hébergé, chaque requête requiert une quantité importante de ressources de calcul et monopolise ces ressources pour un temps important (environ 1min30 par document). Ces coûts en ressources et en temps de calcul couplés à une précision moindre, rendent difficile l'exploitation de modèle LLMs génératif auto-hébergé pour la tâche d'extraction d'information à travers une interface en ligne.

MyoExtract est un outil qui peut permettre d'accélérer le processus de numérisation des comptes rendus de biopsie notamment dans **IMPatienT**. En effet, grâce à cette méthode de détection automatique, il est possible de préremplir les formulaires de numérisation des données dans **IMPatienT** en extrayant automatiquement les données de bases (âge, muscle, numéro de biopsies, anomalies de base...). Cette approche permet un gain de temps important, car il est alors possible d'extraire ces informations d'une masse de rapports sans travail manuel d'annotation, cette méthode apporte une solution aux soucis de mise à l'échelle d'**IMPatienT** dans le cadre de traitement d'une grande quantité de données.

7.3 *MyoClassify* : un outil d'aide au diagnostic

L'outil *MyoClassify* a pour objectif de suggérer un diagnostic parmi les 3 types majoritaires de MC (NM, COM, CNM) de manière automatique sur la base du texte du rapport de biopsie. Pour cela, nous avons utilisé comme jeu de données un corpus élargi de 192 rapports de biopsies fournis par l'institut de myologie de Paris labellisés selon 5 classes (tableau 7.4 : NM, COM, CNM, diagnostic différent des 3 sous-types majoritaires (*non-CM*) et pas de diagnostic final établi (*UNCLEAR*).

7.3.1 Méthodologie

La figure 7.6 représente l'ensemble des étapes réalisées pour préparer les données et entraîner un modèle de classification. Pour l'ensemble de ces rapports, nous avons réalisé une étape de détection de texte par OCR avec *Tesseract* (présenté dans le chapitre 4 "Matériels et méthodes" ainsi que dans le chapitre 5 sur **IMPatienT**). Ensuite, nous avons retiré les conclusions des rapports (indiquant la décision de diagnostic final, c'est-à-dire le label), afin que le modèle

Date d'envoi : 12 juin 2018

INSERM U. 153
BIOLOGIE ET PATHOLOGIE NEURO-MUSCULAIRE
17 RUE DU FER-A-MOULIN
75005 PARIS

COMPTE-RENDU DE BIOPSIE

Nom du patient : Pierre Laroche
né le 07/04/1994, 24 ans
Muscle biopsié : Quadriceps

Date de la biopsie : 07.05.18
Numéro de biopsie : 7377-07

COUPES AU CRYOSTAT DU FRAGMENT CONGELE A -160°C

Hématoxyline-éosine et trichrome de Gomori :

- La biopsie est anormale, les fibres musculaires ont des tailles très inégales avec deux populations de fibres distinctes : une de taille normale et une de petite taille (atrophique).
- La plupart des noyaux sont en situations normales, mais certaines fibres atrophiques ont des noyaux internalisés voir centralisés
- Au trichrome de Gomori, on observe la présence d'agrégats sombres dans les fibres atrophiques. Certaines fibres ont un aspect violacé
- Pas de signe de nécrose ni de régénération.
- Pas d'augmentation du tissu conjonctif interstitiel

Coloration ATP (pH 9.4): Quasi-uniformité de type I. Pas de regroupement des fibres de même type

Coloration PAS : Surcharge importante en glycogène

Coloration Soudan : Pas de surcharge en lipides

CONCLUSIONS :

Atrophie et quasi-uniformité de type I
Présence d'agrégats sombres dans les fibres musculaires
Le diagnostic le plus probable est une myopathie à némaline avec une forte disproportion des types de fibre.

Dr. Louis Dupont

FIGURE 7.5 – **Exemple de compte rendu de biopsie fictif complet**. Comme le vrai rapport de l'institut de Myologie de Paris ce rapport est constitué d'un entête composé d'informations générales, d'un corps de texte composé des observations histologiques pour diverses techniques de coloration et d'une conclusion avec le diagnostic final.

TABLEAU 7.3 – **Résultats de *MyoExtract* pour GPT-3.5-turbo and Vicuna7B**. *GPT-3.5-turbo* extrait l'ensemble des informations demandées de façon satisfaisante, tandis que *Vicuna7B* n'en extrait que 6 sur 13.

Information	**GPT-3.5-turbo**	**Vicuna7B**
Nom	Pierre Laroche	Pierre Laroche
Âge	24	N/A
Date de naissance	07-04-1994	07-04-1994
Date de biopsie	07-05-2018	N/A
Date d'envoi	12-06-2018	N/A
Muscle	Quadriceps	quadriceps
N° Biopsie	7377-07	7377-07
Diagnostic	myopathie à némaline	myopathie à némaline avec forte disproportion des types de fibre
Anomalie PAS	yes	yes
Anomalie Soudan	no	N/A
Anomalie COX	N/A	N/A
Anomalie ATP	quasi-uniformité de type I	N/A
Anomalie Phospho.	N/A	N/A

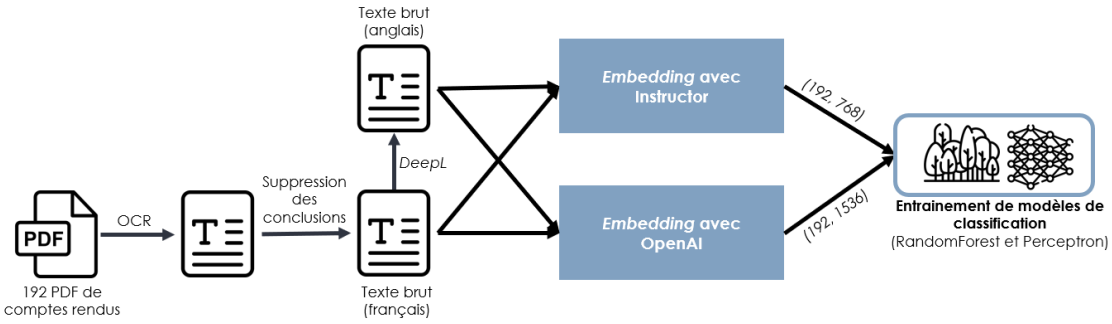


FIGURE 7.6 – **Étapes de préparation et d'entrainement des modèles de *MyoClassify***. Les comptes rendus au format PDF sont convertis en texte dont la conclusion (diagnostic) est supprimée. Ensuite, chaque compte rendu est traduit en anglais. Puis les comptes rendus français et anglais sont convertis en vecteurs numériques par un des modèles d'*embedding* utilisés. Ces vecteurs numériques servent de jeu d'entrainement à un modèle de classification de type forêt aléatoire et perceptrons multicouches.

n'ait pas accès au diagnostic réel pour prédire le diagnostic. Enfin, à partir de ces conclusions, nous avons labellisé à la main chaque rapport avec un diagnostic parmi les 5 catégories listées ci-dessus.

Le contenu de chaque rapport (texte brut sans la partie conclusion) a été traduit en anglais grâce à l'API DeepL afin de comparer les performances sur les textes anglais (traduit) et français (orignaux). Puis ces textes ont été encodés numériquement grâce à deux modèles LLMs d'*embedding* (présenter dans le chapitre 4 "Matériel et méthodes") : le modèle d'OpenAI (disponible uniquement *via* API) et le modèle *Instructor-Large* (auto-hébergé). Les modèles d'*embedding* sont des modèles qui prennent en entrée un texte (un mot, une phrase, un paragraphe ou un document) et qui produisent en sortie un vecteur numérique de grande taille capturant le sens sémantique du document d'entrée. Le modèle d'*embedding* commercial d'OpenAI qui transforme les documents en vecteur de taille (1, 1536), tandis que le modèle auto hébergée libre et open source nommé *Instructor-Large* qui transforme les documents en vecteur de taille (1, 768). Ces modèles sont des boites noires, c'est-à-dire que la signification des centaines (voire des milliers dans le cas d'OpenAI) de valeurs numériques décrivant le document n'est pas connue, cependant elles représentent le sens sémantique du texte. À partir de ces 4 jeux de données (192 rapports dans 4 conditions : français/anglais et *embedding* par OpenAI/Instructor), nous avons entrainé et comparé les performances pour la prédiction de diagnostic de deux algorithmes : les *random forest* et les perceptrons (réseaux de neurones simples). Nous avons retiré du jeu de données les 54 rapports sans diagnostic, car ils ne peuvent pas être utilisés pour l'entraînement des modèles à apprentissage supervisé, ce qui aboutit à 138 rapports utilisés sur 4 labels différents pour l'entraînement des modèles (NM, COM, CNM, non-CM).

Tableau 7.4 – **Nombre de comptes rendus de biopsies par diagnostic**. Au total, ce sont 192 comptes rendus de biopsies répartis sur 5 labels différents, dont les 3 grands sous-types de myopathies congénitales (NM, COM, CNM), un label pour les comptes rendus sans diagnostics (UNCLEAR) et un label pour les comptes rendus non liés aux myopathies congénitales.

Diagnostic	**Nombre de rapports**
Myopathie à Némaline (NM)	44
Myopathie à Cores (COM)	48
Myopathie centronucléaire (CNM)	16
Diagnostic non établi (UNCLEAR)	54
Autre (non-CM)	30

7.3.2 Résultats des entraînements et performances des systèmes d'*embedding*

Au total, 8 conditions expérimentales pour la prédiction de diagnostics ont été évaluées : *Embedding* OpenAI vs Instructor, rapports en Français vs traduits Anglais, *Random Forest* vs Perceptrons (représenté en figure 7.6). Pour chacune des conditions expérimentales, les hyperparamètres des modèles ont été optimisés par grille et les performances ont été évaluées grâce à 10 cross-validations. Ceci a été fait pour (i) obtenir des modèles avec les meilleures performances possibles (optimisation par grille) et (ii) avoir une estimation robuste des performances (moyenne sur 10 essais par cross-validation). L'ensemble des résultats de ces entraînements (métriques de performances et modèles) sont disponibles en ligne à l'adresse : https://wandb.ai/lambda-science/myo-text-classify/reports/MyoClassify-all-conditions-results--Vmlldzo0NDMyMTcw.

Sur le tableau 7.5 et les figures 7.7 et 7.8 on observe que, globalement, les performances à travers les conditions en termes de score F1 sont situées dans un intervalle entre 0.51 et 0.68 et

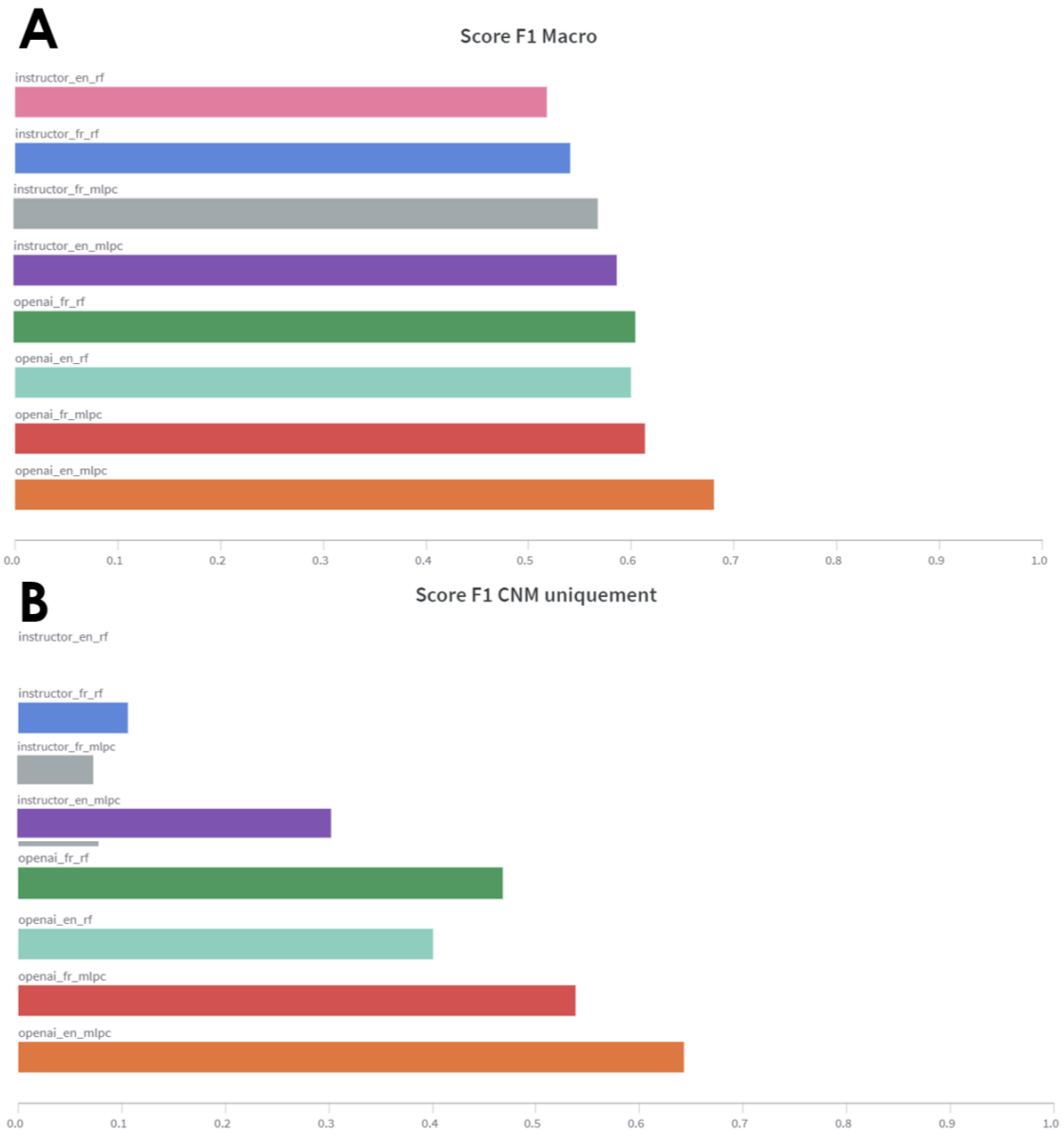


FIGURE 7.7 – **Histogrammes des performances des modèles *MyoClassify*. (A)** Pour le score F1-macro **(B)** Pour le score F1 pour la classe minoritaire (CNM) uniquement. Globalement, le modèle le plus performant est le modèle *OpenAI_EN_MLPC* pour les deux métriques présentées.

TABLEAU 7.5 – **Récapitulatif des performances des modèles *MyoClassify***. Pour chaque modèle, les valeurs d'exactitude, d'exactitude équilibrée, de score F1 pondéré et macro et de score F1 uniquement pour le classe CNM ont été mesurées.

Nom	Exactitude	Exact. Equi.	F1 pond.	F1-Macro	F1 CNM
Instructor FR RF	0.65	0.56	0.62	0.54	0.11
Instructor EN RF	0.67	0.54	0.62	0.52	0.00
Instructor FR MLPC	0.67	0.57	0.66	0.56	0.08
Instructor EN MLPC	0.64	0.58	0.64	0.58	0.32
Openai FR RF	0.61	0.56	0.60	0.58	0.45
Openai EN RF	0.65	0.5792	0.64	0.60	0.40
Openai FR MLPC	0.64	0.6	0.64	0.61	0.54
Openai EN MLPC	**0.70**	**0.67**	**0.69**	**0.68**	**0.64**

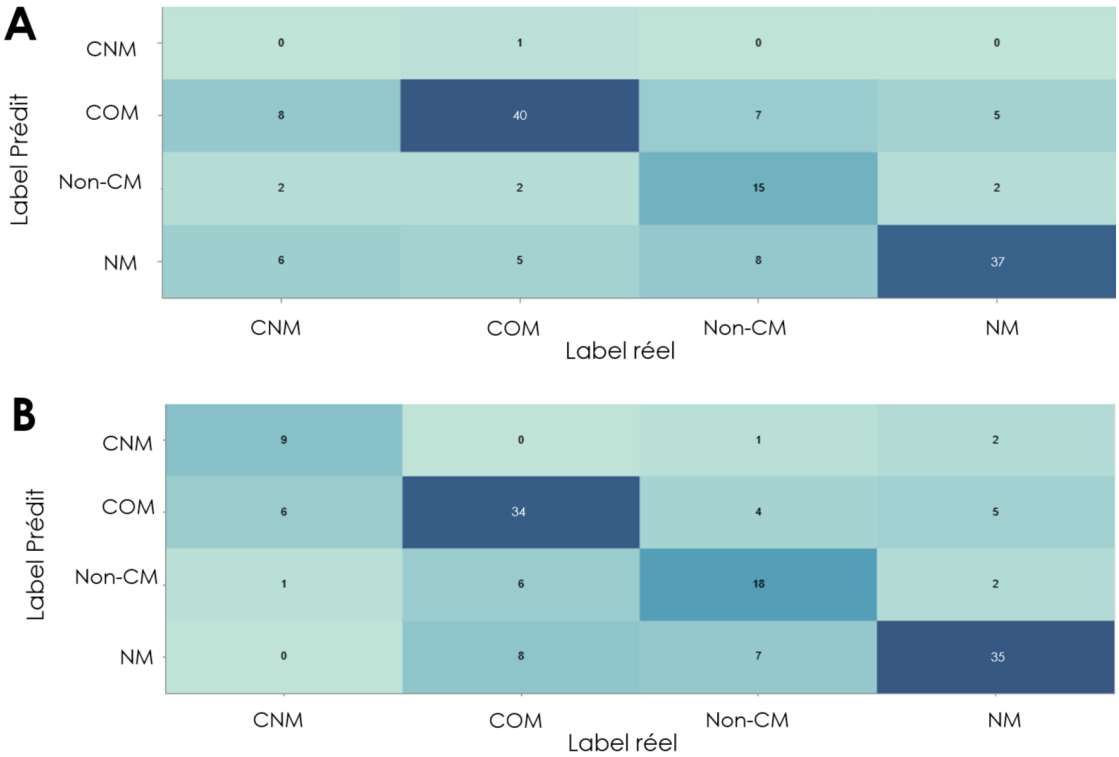


FIGURE 7.8 – **Matrice de confusion des modèles *MyoClassify***. **(A)** Pour le moins bon (*instructor_en_rf*) **(B)** pour le meilleur modèle (*openai_en_mlpc*) en termes de score F1. Le moins bon modèle obtient de meilleurs résultats que le meilleur modèle sur les classes majoritaires (COM et NM) mais de très mauvais résultats sur la classe minoritaire (CNM, 0 bonne classification contre 9).

que donc toutes les méthodes ont réalisé des erreurs de prédiction. Les modèles entraînés sur la base du modèle d'*embedding* auto-hébergé *Instructor* ont de moins bonnes performances globales à travers toutes les conditions. Cependant si l'on observe la matrice de confusion (7.8), on observe que le moins bon modèle *Instructor* a obtenu de meilleures performances de classifications que le meilleur modèle *OpenAI* pour les deux classes de myopathies majoritaires : NM et COM (40 prédictions correctes contre 34 pour les NM, et 37 prédictions correctes contre 35 pour les COM). Cela indique donc que les métriques de performances globales sont donc très influencées par les performances du modèle sur les CNM.

Pour la classe minoritaire (les CNM avec 16 rapports), les performances des modèles sont très faibles pour l'*embedding* du modèle *Instructor*. Par exemple, pour notre modèle *Instructor_FR_RF*, aucune CNM n'a été prédite (matrice de confusion 7.8). Le modèle *OpenAI_EN_MLPC* quant à lui obtient de meilleures performances et a été capable de prédire 9 des 16 CNM.

Dans le cadre du modèle d'*embedding* d'OpenAI, la traduction des rapports en anglais a permis d'obtenir de meilleures performances dans toutes les conditions. De même, l'utilisation d'un perceptron multi-couche a permis d'obtenir de meilleures performances en termes de score F1 dans toutes les conditions par rapport à la *random forest*.

En termes de performances brutes, il semble recommandable de : (i) traduire les comptes rendus en anglais (ii) d'utiliser le modèle d'OpenAI pour l'*embedding* et (iii) d'entrainer un perceptron multi-couche pour apprendre à différencier les diagnostics en fonction de la représentation numérique des rapports (*embeddings*).

7.4 *MyoSearch* : un moteur de recherche de patients

En utilisant les techniques d'*embedding* pour représenter un texte sous forme numérique en capturant son sens sémantique, il est alors possible de calculer un score de similarité entre une requête en texte libre et une masse de documents pour trouver le document le plus proche. Avec *MyoSearch* nous avons créé un outil qui permet de faire des requêtes en texte libre parmi l'ensemble des rapports de biopsies de patients. Il est alors possible de chercher rapidement chez quels patients un symptôme ou diagnostic particulier est présent. Cette création de moteur de recherche est totalement automatique et ne nécessite aucun travail d'annotation, contrairement à l'annotation de comptes rendus dans **IMPatienT**. Elle se déroule en deux phases : (i) l'intégration des données pour constituer la base de données puis (ii) la phase de requêtage de la base de données en fonction de l'entrée de l'utilisateur.

7.4.1 Intégration des rapports : création de la base de données de vecteurs

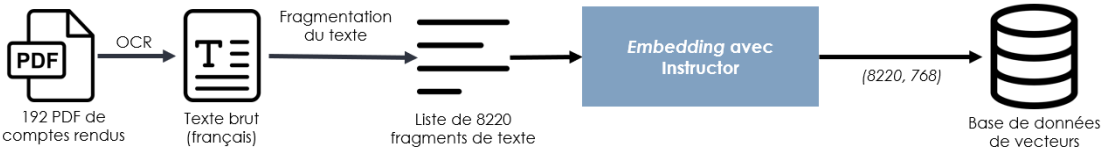


FIGURE 7.9 – **Schéma de l'intégration des données pour le moteur de recherche *MyoSearch*.** L'ensemble des comptes rendus de biopsie sont fragmentés et chaque fragment est transformé en vecteur numérique par le modèle d'*embedding*. Chacune de ces représentations est intégrée dans la base de données de vecteurs.

À la différence de *MyoClassify*, cette fois-ci nous ne voulons pas générer 1 *embedding* par document, mais plutôt séparer les documents en fragments et avoir un *embedding* par fragment de document. Nous avons choisi de découper les documents en fragments de la taille d'une phrase et ainsi d'obtenir un *embedding* pour chaque phrase du document. Ceci permet d'obtenir de meilleur résultat lors du requêtage de la base de données, car le sens sémantique de chaque phrase va pouvoir être comparé à la requête, plutôt que la moyenne de l'ensemble du document.

La figure 7.9 présente la phase d'intégration des données. Nous avons d'abord détecté le texte des rapports PDF par OCR. Comme cette détection est hétérogène et bruitée (images de texte dégradées, difficulté de reconnaissance des caractères), il est difficile de trouver les bornes exactes des phrases. Dès lors, nous avons à fragmenter le contenu en morceaux de taille maximale de 100 *tokens* (environ 15 à 30 mots français) avec un recouvrement de 50, soit la taille moyenne d'une phrase en français. Pour 192 rapports, cela représente 8220 fragments de texte. Pour ces 8220 fragments, nous avons calculé leur représentation numérique (*embedding*) et les avons intégrés dans une base de données de vecteurs. Enfin pour chaque fragment, il est possible d'ajouter des métadonnées qui peuvent servir de filtre pour les requêtes comme le diagnostic final ou le gène responsable de la maladie.

7.4.2 Requêtage des données

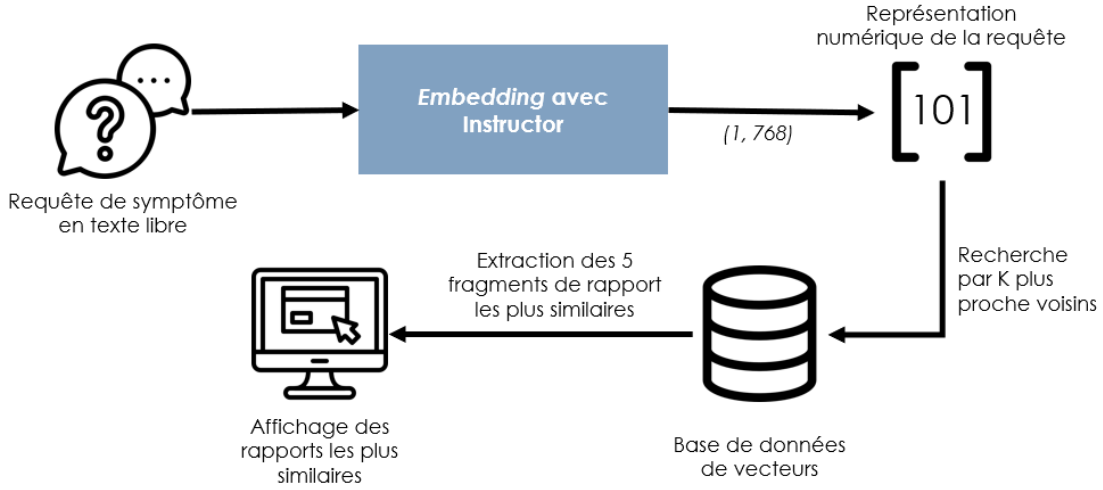


FIGURE 7.10 – **Schéma du requêtage des données dans *MyoSearch***. La requête formulée par l'utilisateur est convertie en vecteur numérique qui est ensuite comparé (par similarité) à l'ensemble des fragments de comptes rendus intégrés dans la base de données de vecteur.

Quand l'ensemble des documents a été découpé et intégré dans la base de données, il est possible de réaliser des requêtes. La figure 7.10 présente la phase de requêtage des données. L'utilisateur peut fournir *via* l'interface web un symptôme d'intérêt en texte libre tel que "surcharge lipidique". Cette requête va ensuite être transformée en vecteur numérique par le modèle d'*embedding Instructor*. Ce vecteur va être comparé à la base de données de vecteurs pour rechercher les tops 5 plus proches voisins grâce à un algorithme nommé *Hierarchical Navigable Small Worlds*, *HNSW*. Les cinq fragments avec les scores de similarité les plus élevés sont ensuite affichés sur l'interface web. Le tableau 7.6 présente les résultats obtenus pour une requête dans MyoSearch. Par exemple pour la requête "surcharge lipidique" les trois rapports les plus proches

font mention d'une surcharge en lipides chez des patients dont (i) le diagnostic n'est pas connu (ii) le diagnostic n'est pas une myopathie congénitale et (iii) chez un patient avec une NM. Cette recherche est aussi multilingue, le modèle d'*embedding* autorise des recherches entre une base de données française avec requête en anglais, ou inversement.

TABLEAU 7.6 – **Exemple d'une requête et des résultats de *MyoSearch*.**

Requête	**Fragment le plus similaire**	**Rang**	**Rapport et diagnostic**
"surcharge lipidique"	"d'inclusions. Il est à noter que l'on observe également une surcharge importante en lipides dans"	1	13405-105.txt UNCLEAR
	"une myopathie congénitale. D'autre part, une surcharge importante en lipides qui nécessite"	2	11391-79.txt non-CM
	"de surcharge en lipides - Technique de Koëlle : CONCLUSIONS : Anomalies caractéristiques d'une"	3	5060-35.txt NM

7.5 Déploiement de l'outil

Développé de façon *open-source*, le code source de **NLMyo** est disponible sur GitHub à l'adresse : https://github.com/lambda-science/NLMyo sous une licence AGPL-3 assurant le statut open source de l'outil. Une version de démonstration en ligne est déployée grâce à *Streamlit* à l'adresse https://lbgi.fr/NLMyo/. Comme **NLMyo** propose l'utilisation de LLMs auto-hébergé, l'outil est hébergé sur un serveur avec un processeur 64 cœurs pour accélérer l'inférence du notre LLMs. Si l'outil n'utilise que l'API OpenAI comme modèle génératif et d'*embedding* alors il est possible d'héberger l'application sur un serveur nécessitant ainsi très peu de ressource de calcul.

7.6 Discussions et perspectives de développement

NLMyo met à disposition des outils permettant le traitement de façon massive de rapports de comptes rendus médicaux et notamment des rapports de biopsie. Cependant, les défis pour rendre l'outil plus robuste sont multiples.

Le premier défi concerne MyoSearch, le moteur de recherche de données de patients. Bien que la méthode soit fonctionnelle et novatrice, les résultats obtenus ne sont pas tout le temps similaire à la requête. Un travail d'amélioration des méthodes de fragmentation et de requêtage est nécessaire. De plus, il n'est actuellement possible que de chercher un symptôme à la fois, il faudrait créer un système permettant de croiser les résultats des requêtes pour plusieurs symptômes, permettant ainsi de chercher un profil de symptômes complet. De plus, l'ajout de métadonnées supplémentaires aux fragments (informations plus complètes sur le patient telles que le gène muté, l'âge, le muscle de la biopsie) sur les patients permettrait de réaliser des requêtes plus fines pour ne sélectionner, par exemple, que les patients liés à un gène en particulier.

Le second défi majeur concerne la protection de la confidentialité des données de santé. En effet, certains outils, pour obtenir les meilleures performances, reposent sur l'utilisation de LLMs externes *via* l'API OpenAI, ce qui est problématique dans le cadre de données sensibles, même anonymisées. Pour cela, nous avons aussi proposé une alternative avec un modèle auto-hébergé, mais pour l'instant celui-ci sous-performe. Par exemple dans *MyoExtract*, les informations extraites sont incomplètes et dans *MyoClassify* les scores d'exactitude et F1 sont globalement plus faibles, voire très faibles, pour la classe minoritaire (CNM). Cependant, la recherche en terme de LLMs est un domaine très dynamique et il est très probable qu'une solution auto-hébergée et performante soit disponible sous peu.

Chapitre 8

Vers une génération de rapports de biopsie automatique avec MyoQuant

Avec **IMPatienT** et **NLMyo** nous avons développé des outils capables de traiter et d'explorer les comptes-rendus de biopsies textuels, permettant ainsi une approche rétrospective de l'ensemble des patients connus à ce jour. Cependant, il est intéressant d'explorer aussi comment l'analyse d'images par IA peut permettre de générer automatiquement des comptes-rendus de biopsies. Cette approche permettrait un gain de précision et de temps dans l'évaluation des biopsies. Tout d'abord un gain de temps, car une approche par IA permettrait de détecter et de quantifier les marqueurs pathologiques sur des biopsies de grande résolution dans de multiples colorations de manière automatique, libérant ainsi le temps utilisé pour l'interprétation des coupes histologiques. Ensuite un gain de précision, car l'évaluation des biopsies par un expert humain n'est en général que qualitative. La description de phénotype se limite généralement à des adjectifs de quantité tels que "peu", "moyen" ou "beaucoup". Grâce à une approche de comptage par IA, il est alors possible d'obtenir la quantité précise pour chaque marqueur évalué tel que le nombre de fibres présentant un noyau centralisé et ainsi de pouvoir établir des seuils pour une analyse plus approfondie.

Le stockage, la gestion et l'exploitation des images en microscopie de grande taille sont des problématiques pour lesquelles de nombreux outils ont été développés. Dans un effort d'uniformisation des formats et des interfaces, on peut notamment citer la plateforme OMERO (Open Microscopy Environment Allan et al., 2012). Cette plateforme cherche à unifier les scientifiques et les développeurs autour de procédures et de format standard de stockage et d'ajout de métadonnées pour les données expérimentales de microscopie. D'autre initiative telle que Cytomine (Marée et al., 2016) a été développée pour permettre l'analyse collaborative d'images biomédicales. Cependant, encore aujourd'hui, ces outils de gestions des données d'imagerie ne sont pas largement répandus parmi les chercheurs. De plus pour la quantification d'images, il est souvent préféré et nécessaire de développer des scripts d'analyse taillés spécifique au besoin d'analyse du projet. Dans ce cadre-là, nous avons développé une série de méthodes génériques applicables pour l'analyse de biopsie du muscle.

MyoQuant (figure 8.1), l'outil présenté dans ce chapitre, est une suite de méthodes pour quantifier différents marqueurs histopathologiques dans les biopsies de MC. **MyoQuant** intègre à la fois des méthodes algorithmiques simples basées sur des modèles d'IA généralistes en histologie comme CellPose (Stringer et al., 2021) et Stardist (Weigert et al., 2020), soit des

méthodes basées sur des modèles IA que nous avons développés à partir de nos données. Actuellement, **MyoQuant** est capable de quantifier des marqueurs pathologiques dans trois des cinq colorations réalisées en routine lors du diagnostic des MC : la centralisation des noyaux à la coloration HE, un déséquilibre dans le ratio des fibres de type 1 et 2 à la coloration ATPase et une répartition anormale des mitochondries dans les fibres musculaires à la coloration SDH. Dans ce chapitre, nous allons voir comment ont été implémentés ces solutions de quantification automatique et des exemples d'application.

FIGURE 8.1 – **Logo de MyoQuant**. MyoQuant est un outil permettant de quantifier automatiquement des marqueurs pathologiques dans les biopsies musculaires. Cet outil open source est disponible en ligne de commande (**MyoQuant**) et en version de démonstration en ligne (**MyoQuant**-Streamlit). https://github.com/lambda-science/MyoQuant

8.1 Analyse de la position des noyaux cellulaires

Dans un premier temps, nous nous sommes intéressés à l'analyse de la position des noyaux cellulaires dans les fibres musculaires. Dans un muscle sain, les noyaux sont en général en périphérie des fibres, tandis que dans les muscles des patients atteints de MC et particulièrement dans les CNM, les noyaux peuvent être internalisés, voire centralisés. Par exemple, dans la figure 8.2, on observe un nombre important de fibres ayant un noyau cellulaire internalisé (flèche noire). Nous avons dès lors développé une méthode pour compter automatiquement le nombre de fibres présentant un noyau internalisé.

8.1.1 Algorithme de quantification

Pour réaliser la quantification des noyaux centralisés, la première étape consiste à obtenir la position de toutes les fibres musculaires et de tous les noyaux de la coupe (segmentation). Pour cela, nous avons utilisé deux modèles d'IA généralistes développés spécifiquement pour l'analyse de coupes histologiques : Cellpose et Stardist. Cellpose nous a permis de segmenter les fibres musculaires, tandis que Stardist nous a permis de segmenter les noyaux cellulaires. La figure 8.3 présente les résultats de la segmentation de la biopsie présentée en figure 8.2. On observe que globalement toutes les fibres musculaires sont bien segmentées, cependant concernant les noyaux cellulaires, certains sont trop peu colorés pour être reconnus par le modèle. C'est notamment le cas pour quelques noyaux centraux sur la gauche de la biopsie (indiqué par des carrés bleus), ce qui sera problématique lors de l'analyse des noyaux, car ils ne seront pas détectés par Stardist.

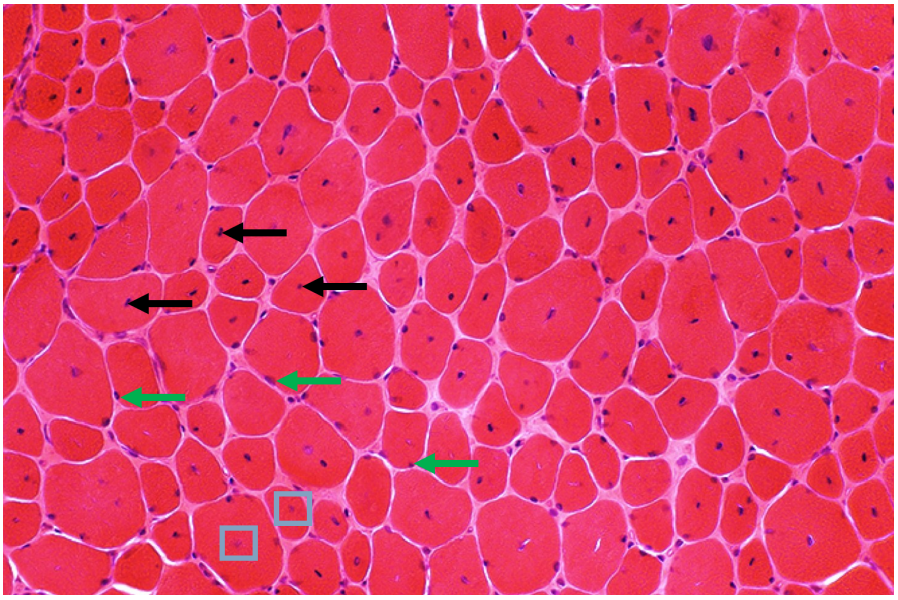


FIGURE 8.2 – **Exemple de biopsie musculaire de CNM à la coloration HE** . On observe des fibres avec des noyaux centralisés (indiqués par des flèches noires) et des noyaux périphériques (flèches vertes). De plus, on observe des noyaux avec une faible intensité de marquage (carrés bleus).

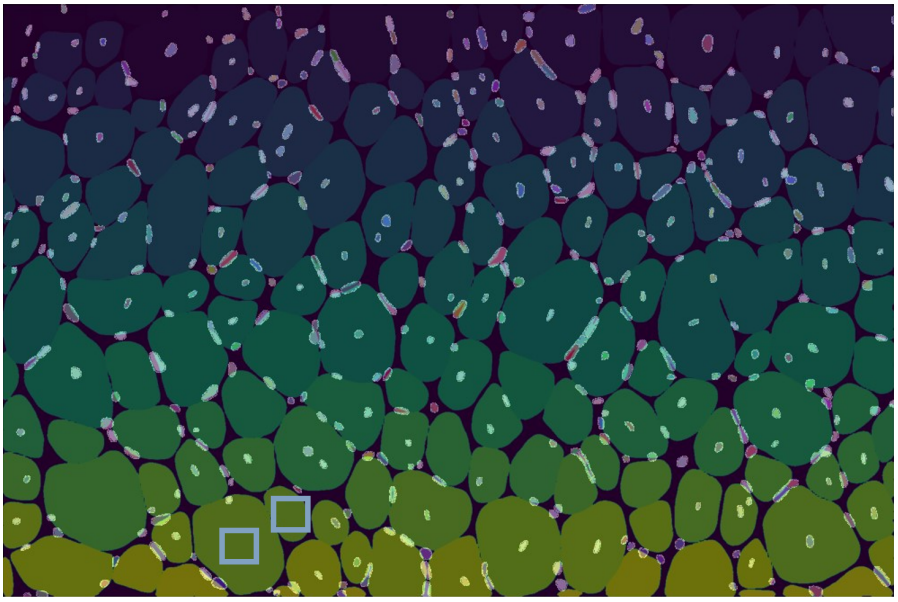


FIGURE 8.3 – **Exemple de segmentation des fibres musculaires et des noyaux cellulaires par Cellpose et Stardist**. Chaque fibre musculaire segmentée est représentée en dégradé de jaune à bleu. Les noyaux détectés par Stardist sont représentés en surbrillance blanche. Les carrés bleus indiquent deux noyaux présents dans la coupe histologique originale, mais non détectés par Stardist.

Après avoir obtenu la position de chaque fibre et noyau, nous évaluons la position de chaque noyau, fibre par fibre. Cette évaluation repose sur le calcul de ce que l'on appelle un score d'excentricité. Ce score est calculé selon la formule suivante :

$$\text{Score d'excentricité} = \frac{\text{Dist. centre fibre et noyau}}{\text{Dist. centre fibre et membrane}}$$

Dans cette formule, la notation "Dist. centre fibre et noyau" représente la distance en pixels entre le centroïde de la fibre musculaire et centroïde du noyau considéré. Et la notation "Dist. centre fibre et membrane" représente la distance entre le centroïde de la fibre musculaire et la membrane cellulaire selon une droite passant par le noyau d'intérêt. La figure 8.4 présente la classification des noyaux d'une fibre musculaire unique. Quatre noyaux ont été détectés dans cette fibre dont trois ont un score d'excentricité supérieur à 0,9 et un inférieur à 0,1. En fixant un seuil de façon empirique à 0,75, on considère alors que cette fibre musculaire possède un noyau internalisé.

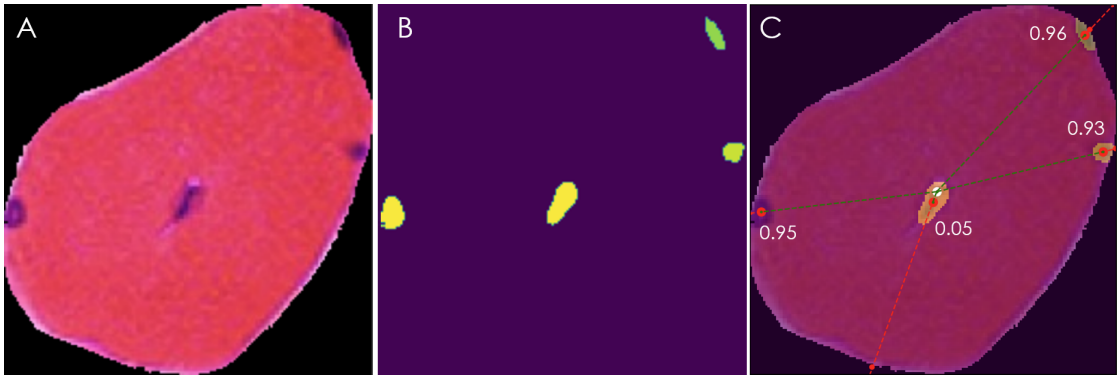


FIGURE 8.4 – **Exemple de classification de la position des noyaux cellulaires d'une fibre musculaire. (A)** La fibre musculaire seule **(B)** masque de segmentation des 4 noyaux pour cette fibre **(C)** schéma de la classification des noyaux avec le score d'excentricité de chaque noyau représentant le ratio de distance : centre de la fibre - noyau versus centre de la fibre - membrane cellulaire.

En comparant l'ensemble des noyaux de chaque fibre au seuil de 0.75, on peut alors quantifier le nombre de fibres musculaires ayant au moins un ou plusieurs noyaux internalisés. Par exemple, pour l'image présentée en figure 8.2, la figure 8.5 présente les résultats de cette classification. Sur cette coupe histologique, on obtient un total de 74 fibres (soit 42% des fibres) avec au moins un noyau internalisé. Cependant, on observe que pour certaines fibres cette détection n'est pas correcte. Par exemple, représenté par un carré bleu et une flèche bleue, on observe trois fibres avec un noyau centralisé, mais considéré comme des fibres avec uniquement des noyaux périphériques. Cela s'explique, car de par leur faible coloration, les noyaux centraux n'ont pas été détectés correctement.

8.1.2 Exemple d'application : quantification de la régénération musculaire

La présence de noyaux centralisés dans les fibres musculaires est un marqueur pathologique dans les biopsies de MC. Cependant, cette centralisation peut aussi être synonyme de régénération musculaire chez les individus sains. Ainsi, la quantification du nombre de noyaux centralisés est donc aussi un moyen de quantifier la régénération musculaire dans une coupe histologique. Dans le cadre d'une collaboration avec l'IGBMC et plus spécifiquement, avec l'équipe Biologie moléculaire et cellulaire des cancers du sein du Dr. Tomasetto, nous avons utilisé **MyoQuant**

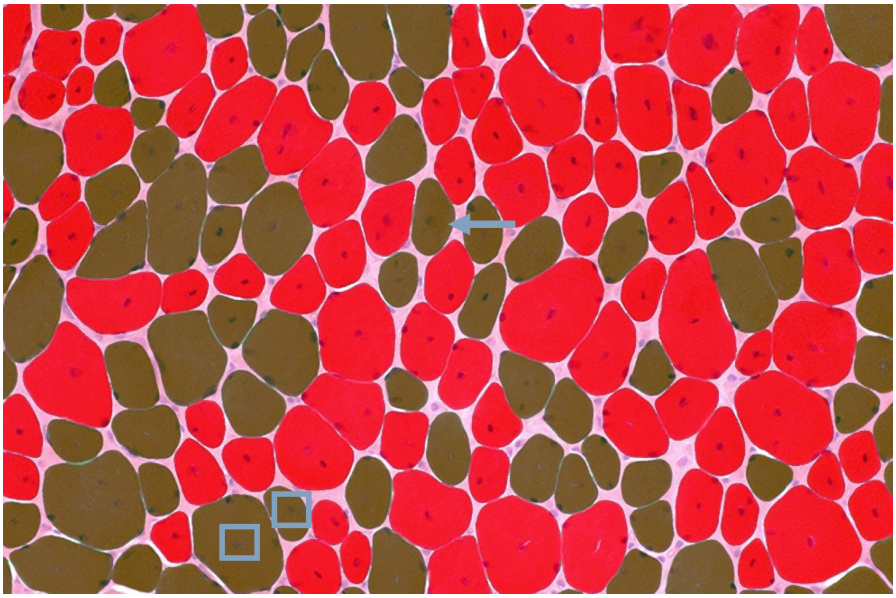


FIGURE 8.5 – **Exemple de classification de biopsie musculaire à la coloration HE.** Colorées en vert les fibres sans noyau internalisé, et en rouge les fibres avec au moins un noyau internalisé (score d'excentricité inférieur à 0.75). Les carrés bleus et la flèche bleue indiquent des fibres dont la classification est fausse (considérées comme n'ayant pas de noyau centralisé).

pour évaluer la quantité de régénération musculaire chez des souris traitées avec une drogue induisant le processus régénératif. Ces images d'histologie sont des images à fluorescence (et non à la coloration HE) avec un fluorochrome pour la membrane cellulaire et un fluorochrome pour les noyaux cellulaires. L'algorithme de **MyoQuant** est directement compatible avec les images à fluorescence et fonctionne de la même façon que pour les images à coloration HE.

La figure 8.6 présente un exemple de coupe complète de biopsie musculaire de souris avec le masque de quantification associé généré par **MyoQuant**. Sur cette image, il y a 6078 fibres musculaires détectées, dont 2285 (environ 37%) sont en régénération. Le tableau 8.1 présente le temps de calcul nécessaire pour chaque étape de la quantification pour la coupe 8.6 et le tableau 8.2 présente les résultats de cette quantification. Une heure de temps de calcul a été nécessaire pour traiter une coupe complète sur une machine classique sans GPU. La majorité de ce temps de calcul a été utilisée par Cellpose pour segmenter les fibres musculaires. Cependant, cette vitesse peut être largement améliorée d'un facteur 5 par l'utilisation de matériel spécifique aux calculs IA (GPU), passant d'environ 1 heure pour Cellpose à moins de 11 minutes. Le temps de calcul de Stardist et de classification des noyaux est négligeable (moins de 1m30). Ainsi, pour une image contenant 6078 fibres et 23 628 noyaux, cela représente environ 1.6 fibre traitée par seconde. Cette quantification a mis en évidence que 37% des fibres présentaient un noyau internalisé et donc sont en régénération.

La figure 8.7 présente l'ensemble des quantifications opérées dans les différentes conditions de traitement et de génotype (au total 35 image de coupe complète (*Whole Slide Image*, *WSI*) analysées). On observe qu'après traitement avec la *Cardiotoxin*, une drogue induisant la régénération musculaire, une proportion significativement supérieure de fibres ayant un noyau internalisé par rapport aux coupes contrôle. Ces résultats confirment que **MyoQuant** est bien capable d'évaluer de façon robuste la présence de noyaux internalisés, un marqueur de la régénération musculaire.

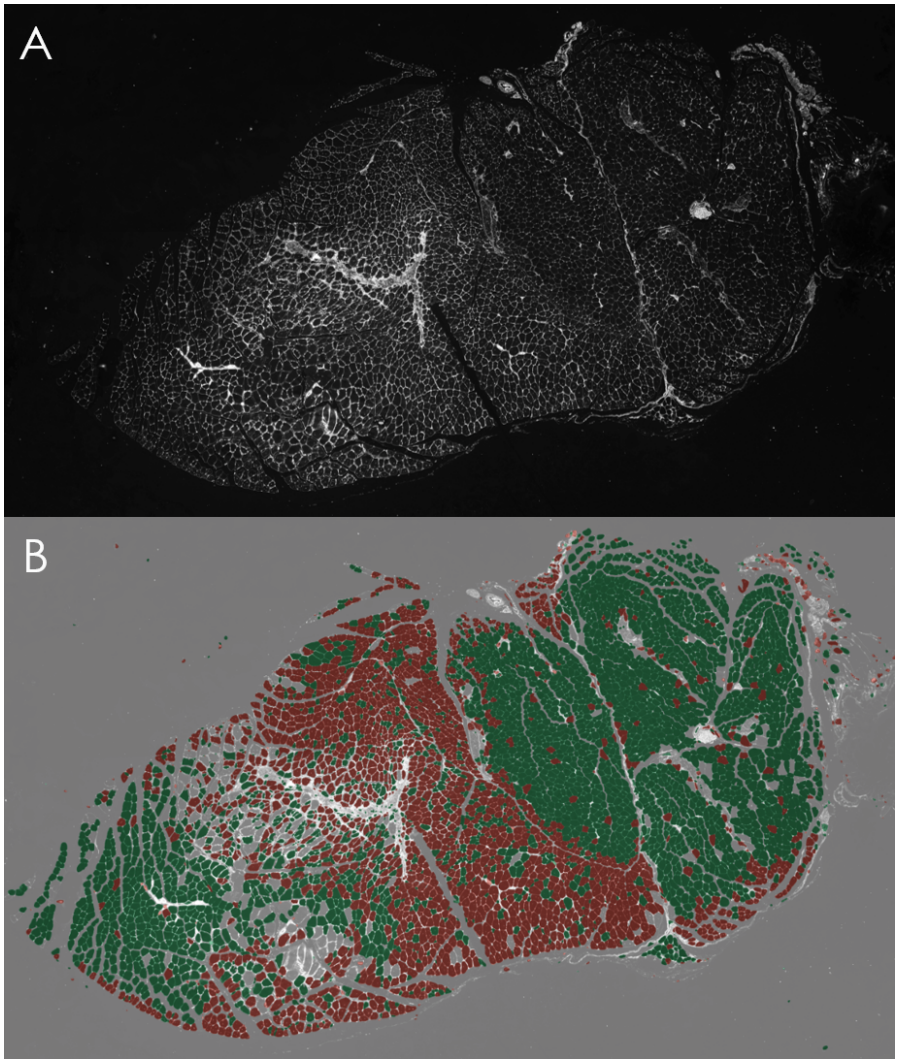


FIGURE 8.6 – **Exemple de classification de biopsie musculaire pour la régénération musculaire. (A)** Image complète de la biopsie musculaire en microscopie à fluorescence. **(B)** Classification des fibres de la biopsie musculaire par l'algorithme d'analyse des centralisations nucléaires. Colorées en vert les fibres sans noyau internalisé (fibres normales), en rouge les fibres avec au moins un noyau internalisé (en régénération)

Tableau 8.1 – **Temps de calcul pour l'analyse des noyaux d'une coupe complète à fluorescence** (6078 fibres, 12000 x 9600 pixels). La classification complète d'une coupe dure environ 1h sur CPU contre 12 minutes sur GPU soit une accélération d'un facteur 5.

Étape	**Temps sur GPU**	**Temps sur CPU (s)**	**Fibres par seconde (sur CPU)**
Cellpose	652	3 782	1.6
StarDist	*mémoire insuffisante*	21	29
Classification des noyaux	68	68	89
Total	**>720**	**3 871**	**1.57**

Tableau 8.2 – **Résultats de la quantification des noyaux d'une coupe complète à fluorescence** (6078 fibres, 12000 x 9600 pixels). Au total, 37% des 6078 fibres ont été détectées comme ayant au moins un noyau centralisé (donc en régénération dans ce cas de figure).

Type	**Valeur**	**Proportion (%)**
N° Fibres	6 078	100
N° Fibres avec 1+ noyau internalisé	2 264	37
N° Noyaux	23 628	100
N° Noyaux internalisés	3 933	16
N° Noyaux périphériques	17 918	76
N° Noyaux non classés (hors fibres)	1 777	8

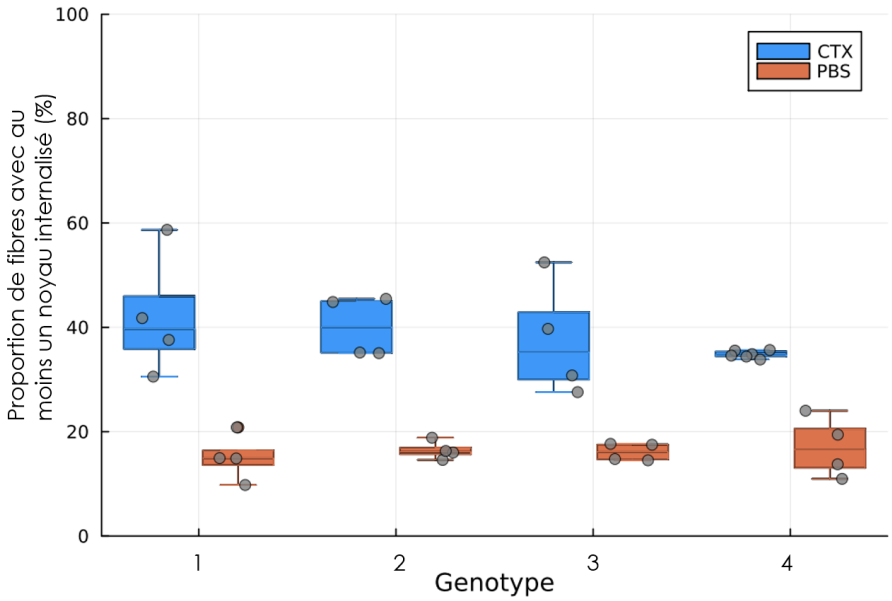


Figure 8.7 – **Résultat de la quantification de la régénération musculaire chez des souris pour 4 génotypes différents** . En bleu sont représentées les souris traitées (bleu) avec une drogue induisant la régénération musculaire (*Cardiotoxin*, CTX). En orange sont représentées les souris de contrôle (solution saline de tampon, PBS).

8.2 Ratio de fibre de type 1 et 2 : classification basée sur l'intensité de coloration

Dans un second temps, nous nous sommes intéressés à l'analyse du ratio des différents types de fibres musculaires dans les biopsies. Dans certaines MC, l'équilibre entre fibres de type 1 (fibre à contraction lente et endurante) et type 2 (fibre à contraction rapide) peut être modifié avec une prédominance des fibres de type 1. Ces deux types de fibres musculaires ont une intensité de coloration différente à la coloration ATPase. À un pH 4.3, les fibres de type 1 sont sombres et les fibres de type 2 sont pâles, et inversement au pH 9.4. La figure 8.8 représente une biopsie musculaire colorée à l'ATPase pH 9.4. On observe assez bien la présence des deux populations de fibres à intensité de coloration distincte et nous avons mis à profit ces différences d'intensité pour développer une méthode de comptage automatique du nombre de fibres de chaque type.

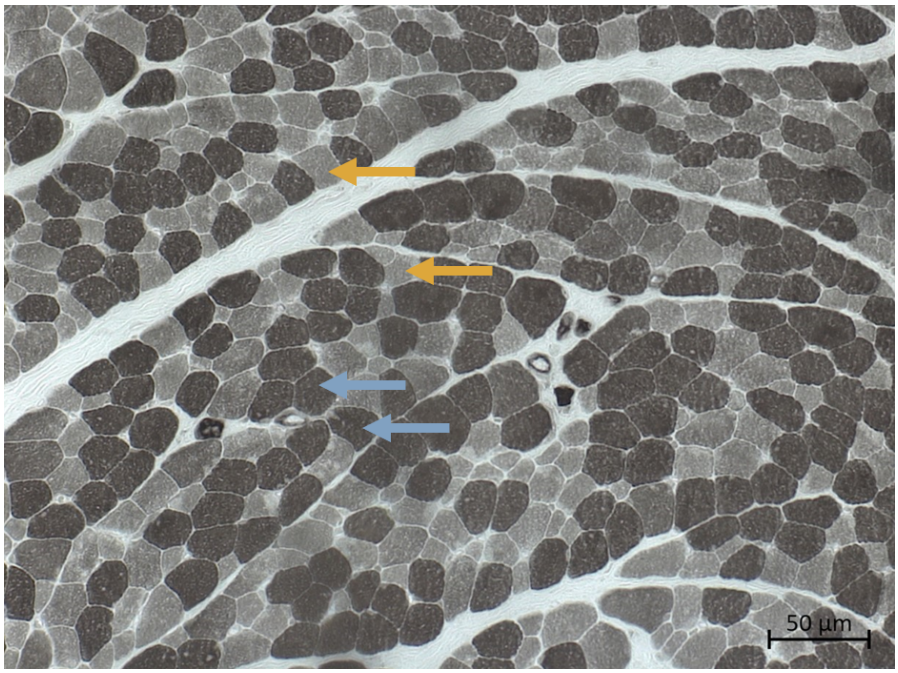


FIGURE 8.8 – **Exemple de biopsie musculaire à la coloration ATPase pH 9.4**. Cette coloration permet de différencier les fibres de type 1 (flèches jaunes) aux fibres de type 2 (flèches bleues).

8.2.1 Algorithme de quantification

Pour réaliser cette quantification, la première étape consiste à segmenter, c'est-à-dire obtenir la position de toutes les fibres musculaires. Comme précédemment, nous avons utilisé Cellpose afin de segmenter les fibres musculaires. Ensuite, pour chaque fibre, nous avons extrait l'intensité moyenne de la fibre et avons réalisé un histogramme. La figure 8.9 présente l'histogramme issu de l'analyse de l'image d'exemple pour la coloration ATPase pH 9.4 (8.8). Le but de la procédure est de déterminer automatiquement les pics présents dans l'histogramme et de trouver les minimums locaux entre les pics pour fixer un ou plusieurs seuils d'intensité. Pour cela, à partir

des valeurs de cet histogramme, une courbe de densité est créée (par méthode de *kernel density estimation (KDE)*). Puis à partir de cette courbe de densité, une méthode de mélange gaussien est utilisée pour déterminer la position des pics dans la courbe de densité. Finalement, le seuil est déterminé automatiquement par notre méthode en trouvant le minimum local de la courbe de densité entre les deux pics obtenus par mélange gaussien.

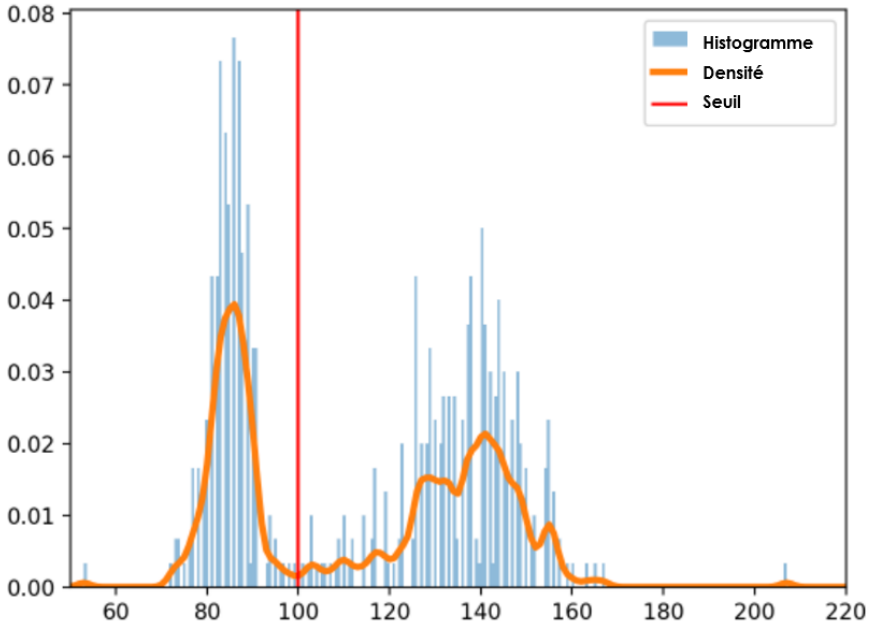


FIGURE 8.9 – **Exemple d'histogramme et courbe de densité biopsie ATPase**. L'histogramme en bleu représente les valeurs moyennes d'intensité de chaque cellule détectée. La courbe en orange est la courbe de densité représentant la distribution des intensités de coloration des cellules. La barre verticale rouge représente le seuil d'intensité défini automatiquement par l'algorithme pour une classification des fibres en deux classes.

À partir de ce seuil, il est possible de classer les fibres musculaires en deux catégories : celles avec une intensité moyenne inférieure au seuil et celles avec une intensité moyenne supérieure. La figure 8.10 présente les résultats de la quantification automatique des fibres de l'image présentée en exemple précédemment. Sur cette image, on a pu quantifier au total la présence de 496 fibres, dont 269 (54%) fibres de type 1 et 227 (46%) fibres de type 2.

8.2.2 Exemple d'application : classification d'une coupe complète avec trois types de fibres

La coloration ATPase peut révéler plus de deux types de fibres musculaires. En effet, les fibres de type 2 ont plusieurs sous-types visualisables dans certaines conditions de coloration. Dans cet exemple, nous avons utilisé la méthode de classification développée pour détecter trois types de fibres. La méthode que nous avons développée est capable d'établir autant de seuils automatiquement que spécifiés par l'utilisateur (et donc de classes). La figure 8.11 présente les résultats de classification d'une WSI de biopsie musculaire colorée à l'ATP pH 4.6. Sur cette coupe, on observe trois populations de fibres : des fibres pâles (fibres de type 2A, flèche noire),

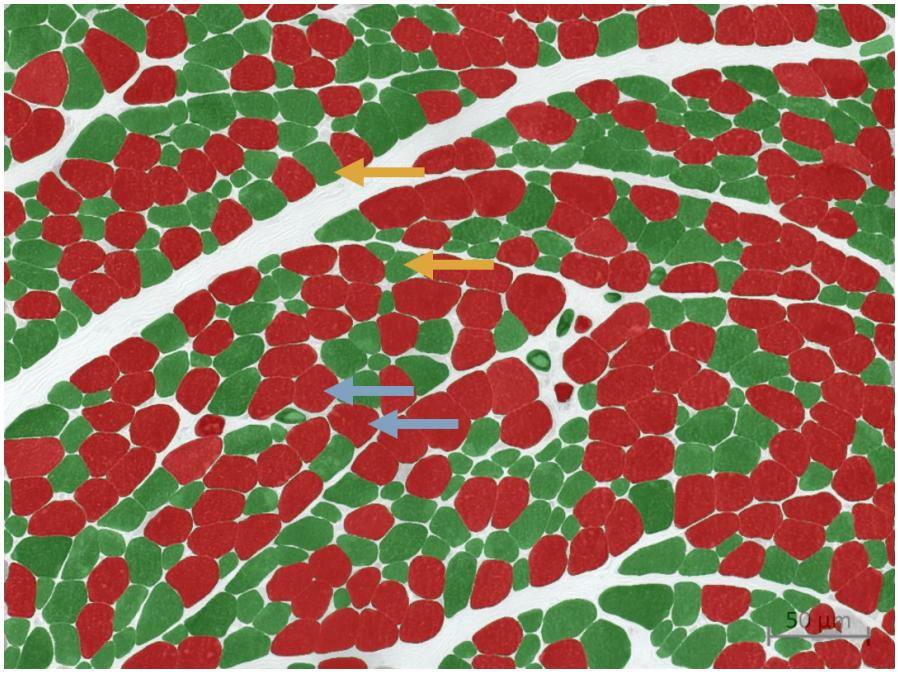


FIGURE 8.10 – **Exemple de classification de biopsie musculaire à la coloration ATPase pH 9.4.** Les fibres ayant une intensité inférieure au seuil (type 2, flèches bleues) sont colorées en rouge. Les fibres ayant une intensité supérieure au seuil (type 1, flèches jaunes) sont colorées en vert.

des fibres intermédiaires (fibres de type 2B, flèche blanche) et de petites fibres très sombres localisées en haut de la coupe (fibres de type 1, flèche verte). La méthode de quantification a alors pu définir deux seuils pour séparer ces trois classes.

Les résultats de cette classification sont référencés dans le tableau 8.4, le temps de calcul et la vitesse de classification sont disponibles dans le tableau 8.3. Au total, 1840 fibres ont été classifiées en 131 secondes (soit 14 fibres par seconde). Il y a une majorité de fibres de type 2A (894, 49%), de fibres de type de 2B (829, 45%) et une faible proportion de petites fibres de type 1 (117, 6%). Les résultats de cette quantification sont visuellement satisfaisants, bien qu'une partie de la biopsie en bas à droite, est repliée sur elle-même et donc apparait avec une intensité de coloration forte. Cette zone a donc été considérée à tort comme des fibres de type 1.

Ces résultats montrent qu'il est nécessaire d'améliorer la robustesse de la méthode de quantification pour prendre en compte la variabilité des échantillons biologiques. En plus des repliements qui influent sur l'intensité de coloration, il peut aussi y avoir des artéfacts de congélations (zone sous forme de bulles blanches dans les fibres, visibles sur la partie gauche de la coupe) qui eux aussi faussent le calcul de l'intensité moyenne des fibres. De fait, il peut être nécessaire d'avoir des méthodes de filtrage en amont de la classification pour ne quantifier que les fibres de bonne qualité, ou d'éliminer au préalable les zones très hétérogènes.

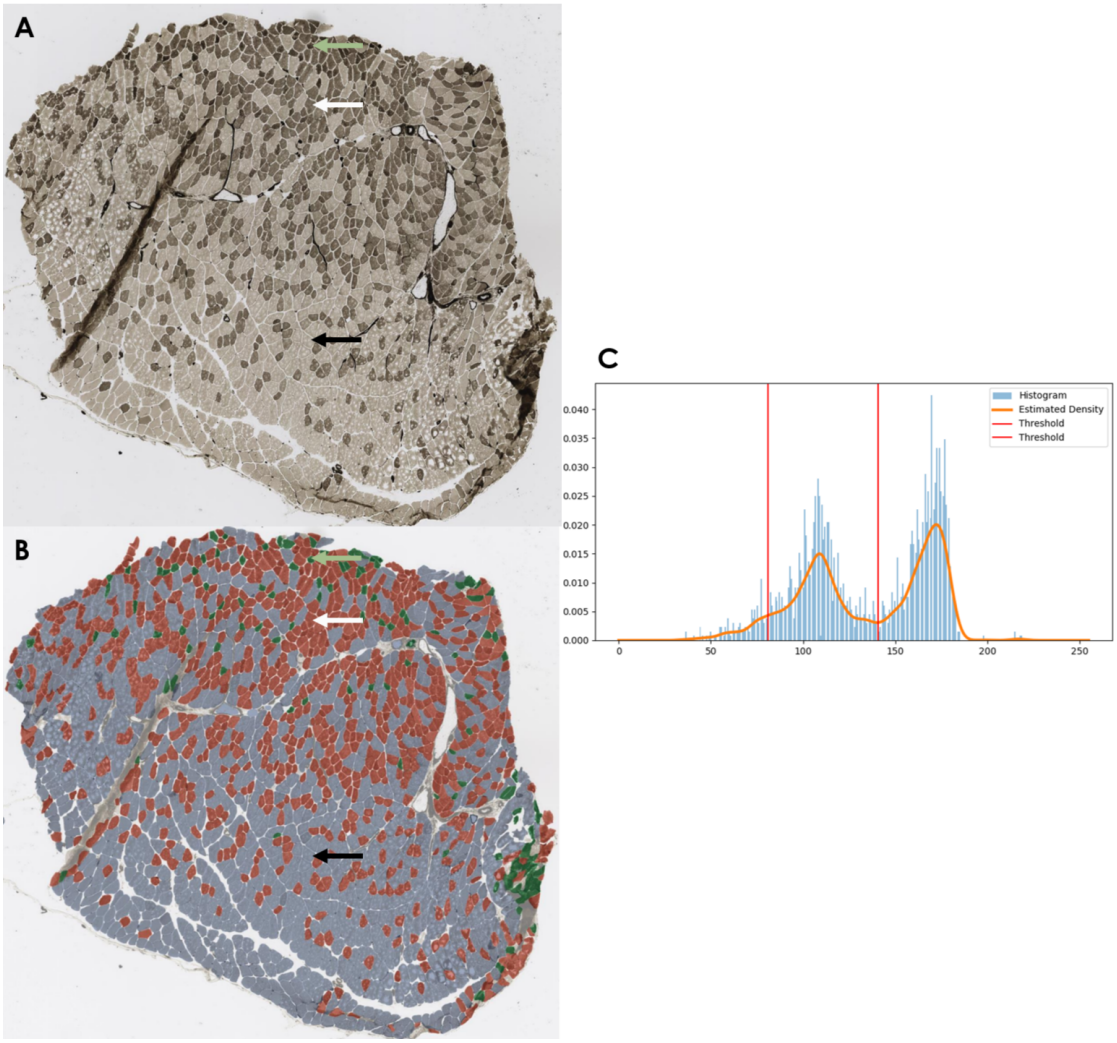


FIGURE 8.11 – **Exemple de classification de biopsie musculaire colorée à l'ATPase pH 4.6 en 3 classes**. **(A)** Image brute de biopsie musculaire ATPase pH 4.6. Pointées par une flèche noire les fibres 2A, blanche les fibres 2B et verte les fibres type 1 **(B)** Image de biopsie dont les fibres ont été colorées en fonction de leur classification : en bleu les fibres de type 2A, en rouge les fibres de type 2B et en vert les fibres de type 1. **(C)** Histogramme et courbe de densité des fibres de la biopsie complète.

TABLEAU 8.3 – **Temps de calcul pour l'analyse des types de fibres d'une coupe complète à la coloration ATPase pH 4.6** (1840 fibres, 6000 x 5600 pixels). La classification complète d'une coupe composée de 1840 fibres sur GPU dure au total 2 minutes et 11 secondes.

Étape	**Temps sur GPU**	**Fibre par seconde (sur GPU)**
Cellpose	113	16
Classification des fibres	18	102
Total	**131**	**14**

TABLEAU 8.4 – **Résultats de quantification des types de fibre d'une coupe complète à la coloration ATPase pH 4.6** (1840 fibres, 6000 x 5600 pixels). Sur cette coupe complète, on obtient un ratio presque égal de fibre 2A et 2B (49% et 45% respectivement) ainsi qu'une très petite proportion de fibres de type 1 (petites fibres fortement colorées, 6%).

Type	**Valeur**	**Proportion (%)**
Fibre type 2A	894	49
Fibre type 2B	829	45
Fibre type 1	117	6

8.3 Répartition des mitochondries : classification par IA

Enfin, nous nous sommes intéressés à l'analyse de la répartition des mitochondries dans les fibres musculaires. Cette répartition peut être anormale dans certaines MC. Cette répartition se visualise grâce à la coloration SDH, révélant l'activité oxydative des fibres musculaires et donc la position des mitochondries. La figure 8.12 présente un exemple de biopsie de muscle d'une souris modèle de myopathie congénitale à la coloration SDH. On observe deux types de fibres, des fibres "normales" (indiqué par des flèches noires) ayant une répartition homogène en coloration et des fibres "anormales" ayant une agglutination de coloration au centre de la fibre (indiqués par des flèches rouges), représentant des agrégats mitochondriaux pathologiques. Dans ce cadre, nous avons développé une méthode capable de détecter et de compter les fibres ayant une répartition mitochondriale anormale, en développant notre propre modèle d'IA.

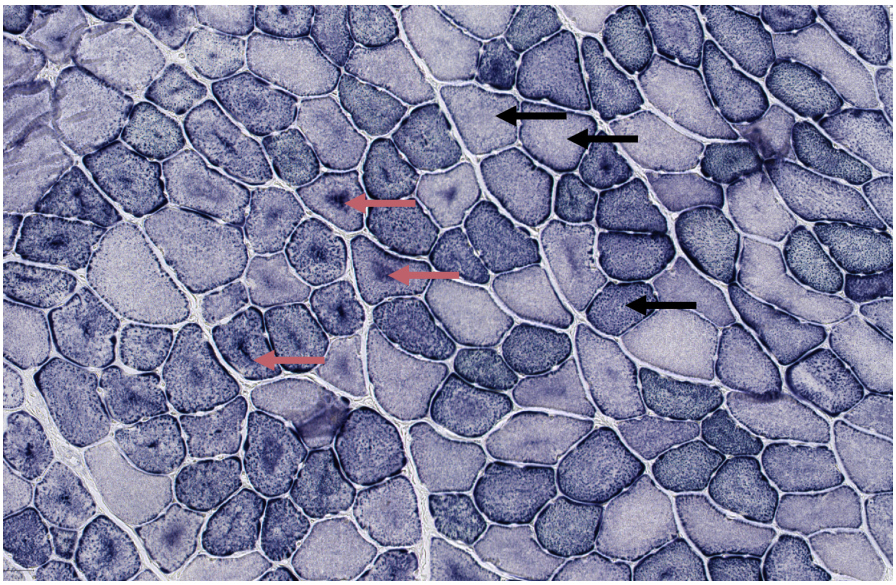


FIGURE 8.12 – **Exemple de biopsie de muscle de souris modèle de CNM à la coloration SDH**. Les flèches noires pointent des fibres ayant une répartition de coloration normale. Les flèches rouges pointent des fibres ayant une coloration anormale avec une tache sombre centrale. Ces fibres ont une répartition mitochondriale pathologique.

8.3.1 Jeu de données d'image de muscle de souris et annotations

Pour constituer le jeu de données nécessaire à l'entrainement de cette IA nous avons utilisés 18 WSI de biopsies musculaires de souris, dont une souris saine (*wild-type*) et 17 WSI de souris modèles de CNM (10 *Bin1-KO* et 7 *Dnm2S619L*). Ces 18 WSI représentent un total de 16 787 fibres musculaires (tableau 8.5). Chacune de ces fibres musculaires a été isolée et extraite de la coupe grâce à Cellpose puis a été annotée à la main en 2 catégories : fibre saine ou anormale, par l'expert ayant généré les images (Quentin Giraud et Charlotte Gineste de l'équipe "Physiopathologie des maladies neuromusculaires" de l'IGBMC). Au total, les 12 730 fibres saines et 4 057 fibres annotées comme anormales ont été séparées équitablement en trois jeux de données : 72% pour le jeu d'entrainement de l'IA, 8% pour le jeu de validation lors de

l'entrainement, et 20% pour le jeu de test des performances de l'IA. Ce jeu de données a été mis à disposition de la communauté scientifique de façon *open-source* sur la plateforme *Hugging-Face* à l'adresse : https://huggingface.co/datasets/corentinm7/MyoQuant-SDH-Data.

Tableau 8.5 – **Répartition des fibres pour le jeu d'entrainement du modèle SDH**. Malgré une répartition avec 1 coupe *wild-type* et 17 coupes de souris modèle de MC, on observe une majorité de fibre saine parmi les 16 787 fibres du jeu d'entrainement (76%).

	Entraînement (72%)	**Validation** (8%)	**Test** (20%)	**Total**
Saine	9 165	1 019	2 546	12 730 (76%)
Anormale	2 920	325	812	4 057 (24%)
Total	12 085	1 344	3 358	16 787

8.3.2 Architecture, entrainement et performance du modèle IA

À partir de ce jeu de données, nous avons entrainé un modèle de CNN. Nous avons sélectionné l'architecture réseau de neurones profonds *Resnet50* pré-entrainée sur *ImageNet*. L'architecture *Resnet50* est une architecture de DNN très utilisée pour la classification d'images et le pré-entrainement sur *ImageNet* permet d'obtenir de meilleures performances et une convergence du modèle accélérée. Pour limiter le sur-apprentissage, nous avons appliqué des techniques d'augmentation de données lors de l'apprentissage par variation de luminosité, contraste, rotation aléatoire, zoom, translation et retournement. De plus, nous avons utilisé une méthode d'arrêt prématuré pour arrêter l'entrainement lorsque les performances sur le jeu de validation ne se sont pas améliorées lors des 10 dernières époques. Enfin, chaque fibre musculaire unique a été redimensionnée à une taille de 256x256 pixels, car le réseau de neurones impose une taille d'image constante.

La figure 8.13 présente les courbes d'apprentissage du modèle SDH. Que ce soit en termes de mesure de la fonction de cout du modèle (*loss function*) ou d'exactitude de classification, on observe qu'après 10 époques (12 minutes d'entrainement), les performances sont maximales pour le jeu de validation et ne s'améliorent plus sur les 10 époques suivantes. Ceci indique qu'après 10 époques, l'apprentissage donne lieu à un sur-apprentissage du jeu d'entrainement. C'est pourquoi grâce à l'arrêt prématuré, nous avons sélectionné le modèle dans l'état après 10 époques comme modèle optimal sans sur-apprentissage.

Après la phase d'apprentissage, pour mesurer les performances du modèle, nous avons utilisé le jeu de test composé de 3358 images non utilisées pour l'entrainement. En comparant les prédictions du modèle sur les images de test à leur annotation par les experts, nous avons obtenu une exactitude de classification de 93,2% et une exactitude pondérée de 91.7%. Autrement dit, le modèle est capable de reproduire l'annotation des deux experts avec une exactitude de 93,2%.

L'ensemble des métriques mesurées lors de l'entrainement sont disponibles de façon open source à l'adresse : https://wandb.ai/lambda-science/myoquant-sdh/reports/Model-Training—Vmlldzo0NDI4MDI4, le modèle ainsi que le code source utilisés pour réaliser cet entrainement de manière reproductible sont aussi disponibles sur la plateforme *HuggingFace* à l'adresse : https://huggingface.co/corentinm7/MyoQuant-SDH-Model.

8.3.3 Exemple d'application

Grâce au modèle développé, il est maintenant possible de classer des fibres individuelles pour détecter les anomalies de répartition mitochondriale. Ainsi pour analyser une image complète,

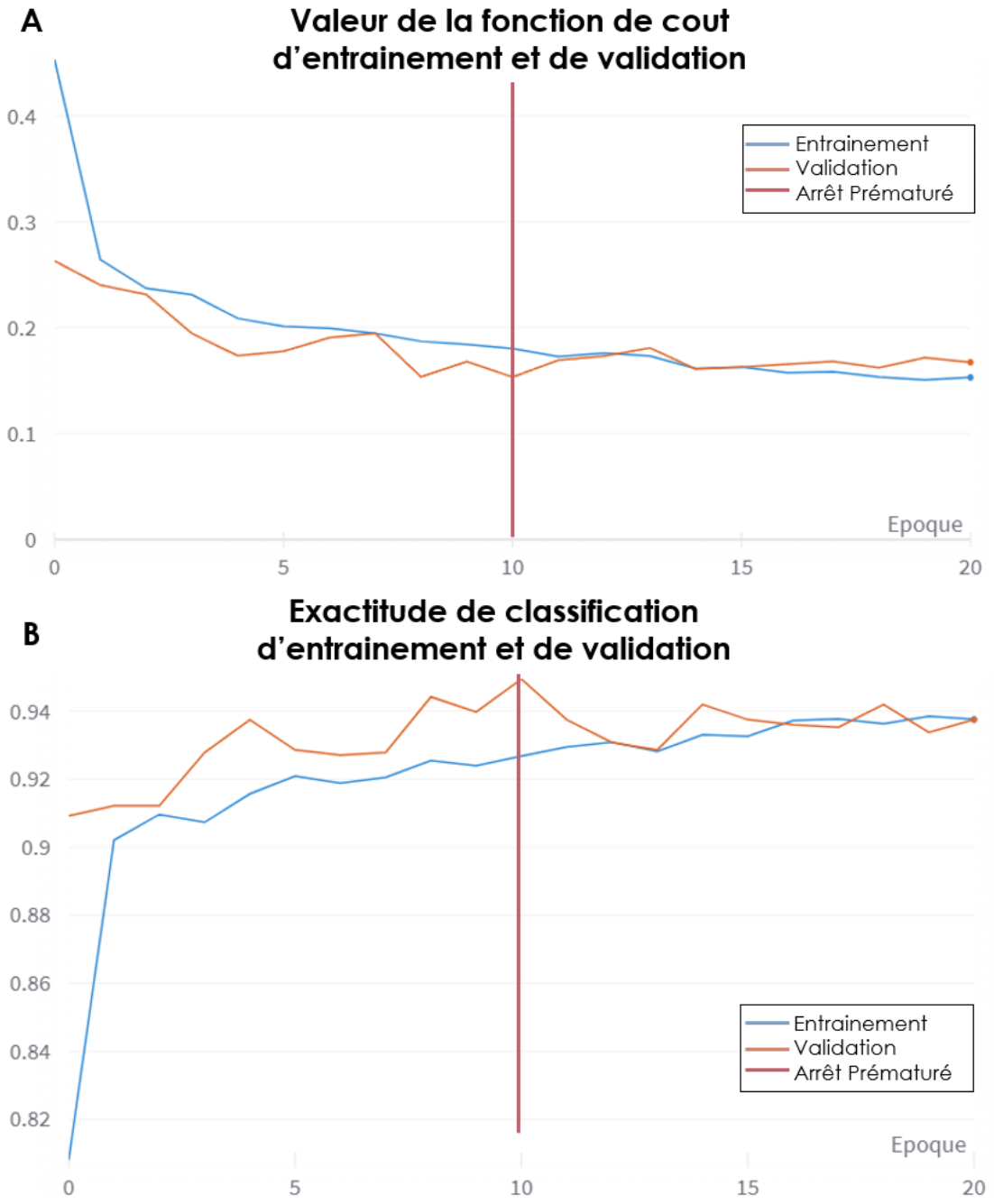


FIGURE 8.13 – **Courbe d'apprentissage du modèle SDH. (A)** La courbe de la fonction de cout **(B)** La courbe d'exactitude de classification pour le jeu d'entrainement (bleu) et le jeu de validation (orange). La barre verticale rouge indique le modèle final sélectionné par le mécanisme d'arrêt prématuré.

on utilise d'abord Cellpose pour segmenter et isoler chaque fibre musculaire de l'image, puis le modèle que nous avons développé pour prédire la classe de la fibre. Par exemple, pour l'image de biopsie présentée ci-dessus (8.12), le résultat de classification est présenté en figure 8.14. Au total, sur 162 fibres détectées, 86 (53%) sont classées comme ayant une répartition mitochondriale anormale et 76 (46%) ont une répartition normale. On observe que ce sont bien les fibres ayant une agglutination de coloration au centre de la fibre (agrégats de mitochondries, sur la gauche de la coupe) qui sont classées comme anormales, confirmant le bon fonctionnement du modèle.

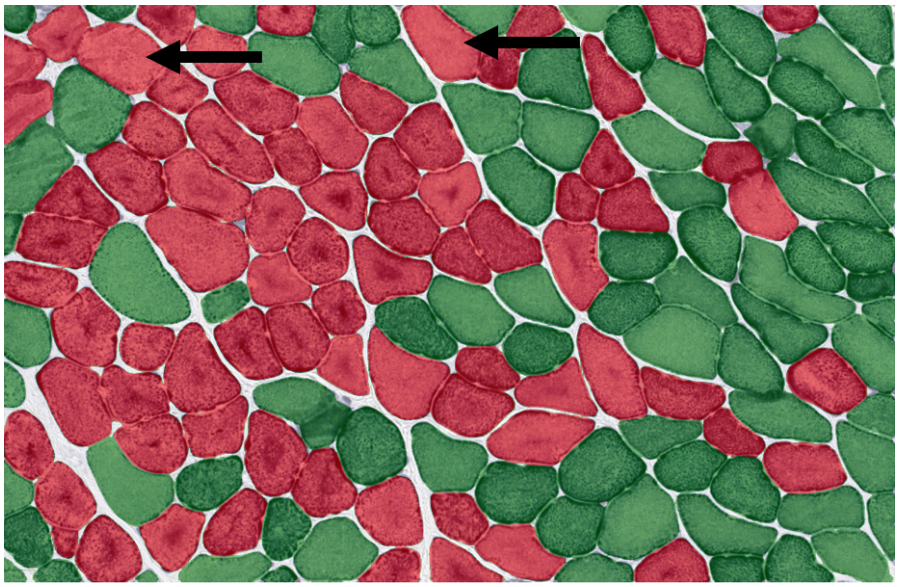


FIGURE 8.14 – **Exemple de classification de biopsie musculaire de souris à la coloration SDH.** Colorées en rouge, les fibres ayant une répartition mitochondriale anormale, en vert, une répartition normale. Les fibres pointées par une flèche noire sont des fibres qui sont classées comme anormales, mais ayant une apparence saine (possible erreur de classification).

Il est aussi possible de classer des WSI de biopsies de muscle complet colorées au SDH. Par exemple, la figure 8.15 présente les résultats de classification d'une coupe complète comptant 2 869 fibres musculaires. En termes de ressources de calcul, l'étape limitante reste le modèle Cellpose qui, pour cette image, requiert trop de mémoire pour fonctionner sur notre matériel (GPU). Ainsi sur CPU, nous avons eu besoin d'environ 40 minutes pour quantifier la coupe, dont la majorité du temps a été utilisé par Cellpose, ce qui représente une vitesse d'environ 1 fibre par seconde. Au final sur la coupe présentée, 2371 fibres sont classées comme saines (83%) et 498 sont classées comme anormales (17%).

8.3.4 Exploration du modèle

Une fois le modèle entrainé, nous avons voulu explorer le modèle grâce à différentes techniques de visualisation pour nous assurer que ses capacités de classification sont bien réelles et ne sont pas dues à un biais lors de l'apprentissage.

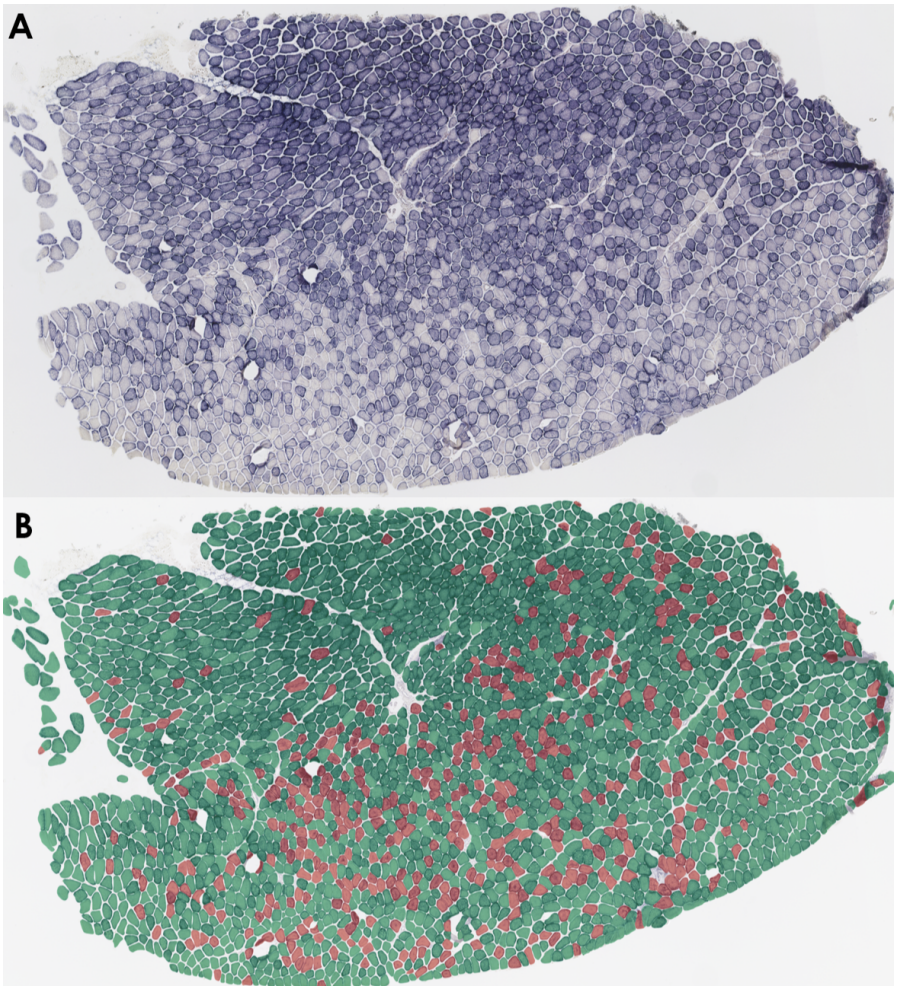


FIGURE 8.15 – **Exemple de classification de coupe complète biopsie musculaire de souris à la coloration SDH. (A)** Biopsie musculaire complète de souris à la coloration SDH. **(B)** Biopsie musculaire complète dont les fibres ont été colorées en fonction de leur classification par le modèle IA : en rouge les fibres ayant répartition mitochondriale anormale, en vert une répartition normale

TABLEAU 8.6 – **Temps de calcul pour l'analyse des fibres d'une coupe complète SDH** (2 869 fibres, 11200 x 6300 pixels). La classification complète de la coupe composée de 2869 fibres a duré 41 minutes sur CPU.

Étape	**Temps sur GPU**	**Temps sur CPU**	**Fibre par seconde (sur CPU)**
Cellpose	*Mémoire insuffisante*	2 407	1,2
Classification des fibres	70	66	43
Total	**>70**	**2473**	**1,15**

TABLEAU 8.7 – **Résultats de quantification des fibres d'une coupe complète SDH** (2 869 fibres, 11200 x 6300 pixels). Sur cette coupe complète, 17% des fibres ont été détectées comme ayant une répartition anormale en mitochondries.

Type	Valeur	Proportion (%)
Fibres saines	2371	83
Fibres anormales	498	17

8.3.4.1 Explicabilité de la classification

Afin de vérifier sur quels critères la classification des fibres est réalisée, nous avons utilisé la méthode Grad-Cam (*Gradient-weighted Class Activation Mapping*, SELVARAJU et al., 2020). Cette méthode permet de générer une carte thermique des images pour visualiser les régions importantes pour la classification selon le modèle. Sur un plan technique, cette carte thermique est réalisée en utilisant les gradients de la sortie par rapport aux cartes de caractéristiques des couches convolutives pour chaque classe, puis une moyenne pondérée est calculée.

La figure 8.16 présente les résultats de cette méthode de visualisation sur 16 fibres musculaires uniques prise au hasard dans le jeu de test. On observe sur cette figure que la zone déterminante (représentée par une couleur rouge) pour la classification des fibres selon le modèle est le centre de la fibre musculaire. Cette observation est cohérente, car nous avons observé que dans la majorité des cas, une fibre est dite anormale lorsqu'elle présente une agglutination de coloration au centre de la fibre (agrégats de mitochondries centralisés). Ce résultat confirme donc que le modèle porte son attention globalement sur la même zone que l'expert lors de la classification manuelle.

8.3.4.2 *Embedding* des images et réduction de dimensionnalité

Pour rappel, la notion d'*embedding* correspond à la transformation d'une donnée en un vecteur numérique de grande taille. Dans **NLMyo** nous avons utilisé des techniques d'*embedding* sur des données textuelles. Ici, nous avons réalisé un *embedding* de nos images, non pas pour faire de la classification, mais pour appliquer des techniques de réduction de dimensionnalité pour visualiser si notre modèle est bien capable de faire une différence nette entre les deux classes.

Pour cela, nous avons utilisé notre modèle SDH et avons retiré la dernière couche de neurones servant à la classification. Ainsi la sortie du modèle correspond à la sortie de la dernière couche convolutive, c'est-à-dire à un vecteur de taille (1, 2048) correspondant aux caractéristiques importantes de l'image extraite par le modèle pour la classification. En réalisant cette opération pour l'ensemble des images, nous obtenons une matrice (16 787, 2048) sur laquelle nous appliquons une méthode de réduction de dimensionnalité (nommée UMAP (MCINNES et al., 2020)), donnant ainsi une matrice (16 787, 2) visualisable facilement en deux dimensions.

Ainsi, la figure 8.17 présente la visualisation obtenue après *embedding* et réduction de dimensionnalité des 16 787 images du jeu de données. Chaque point représente une image du jeu de données, colorée selon son label (annotation) ou sa provenance (modèle de souris). Globalement on observe que sur le premier (et donc le plus important) axe de variance (en abscisse), le modèle extrait des informations permettant de faire la différence entre les fibres saines et anormales. En effet, les deux classes de fibres sont bien séparées avec cependant la présence d'un continuum entre les deux groupes indiquant la présence de fibres avec des profils intermédiaires plus complexes. Le second axe de variance (en ordonnée) indique que le modèle extrait aussi des informations permettant d'établir la provenance de la fibre, c'est-à-dire de quel

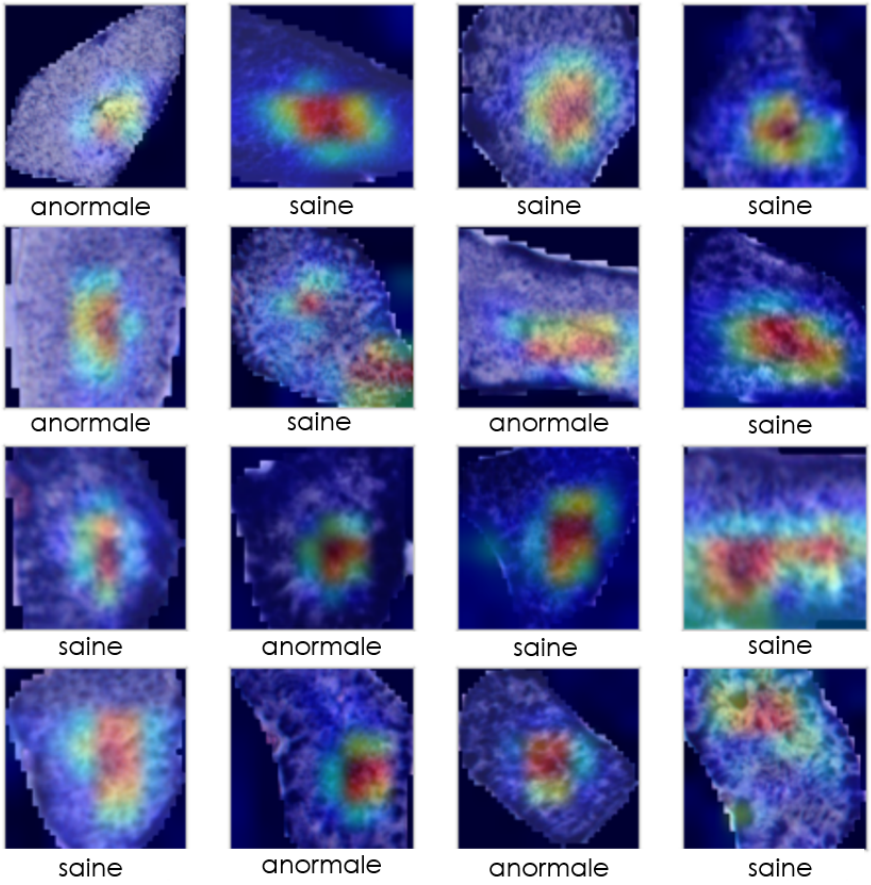


FIGURE 8.16 – **Visualisation par méthode *Grad-Cam* de la classification de 16 fibres par le modèle SDH**. Les 16 fibres ont été prises au hasard du jeu de test. Superposition de l'image originale re-dimensionnée et de la carte thermique générée. La couleur rouge indique une importance forte pour la décision de classification par le modèle. Une couleur bleue indique une importance fiable.

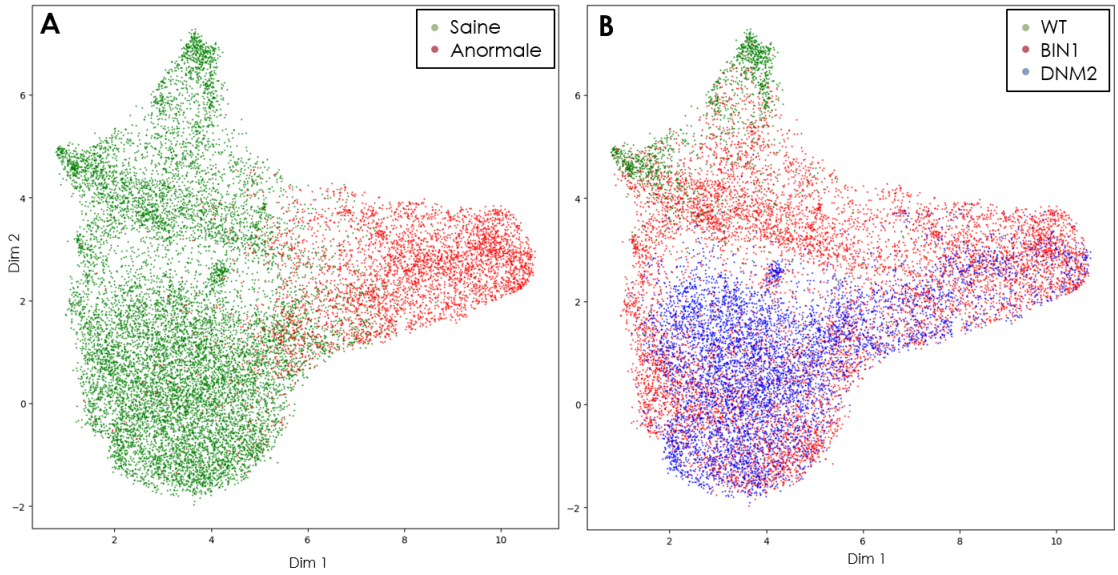


FIGURE 8.17 – **Visualisation de l'*embedding* du modèle SDH des 16 787 images de fibres musculaires après réduction de dimensionnalité par UMAP. (A)** Les fibres sont colorées par label (Saine en vert, Anormale en rouge) **(B)** Les fibres sont colorées par modèle de souris (WT en vert, BIN1 en rouge, DNM2 en bleu).

modèle de souris elle est issue. En effet, on voit un cluster de fibres venant de souris *wild-type* assez compacte en haut de la figure, puis deux clusters à la fois superposés et séparés de fibres provenant de souris modèle *BIN1* et *DNM2*. L'ensemble des ces résultats confirment que notre modèle extrait et base sa classification des fibres sur les caractéristiques intrinsèques des fibres saines ou anormales et non, sur des biais extérieurs.

8.3.4.3 Identification des erreurs d'annotation par le modèle

Lors de la phase de test, le modèle a obtenu une exactitude de classification de 93.2%, il existe donc une marge de progression pour le modèle. Les 6.8% d'erreur peuvent être expliqués par plusieurs raisons. La première est que le modèle n'est peut-être pas assez complexe pour capturer l'ensemble des critères de classification des images et donc a des difficultés pour certains cas complexes. Cependant, cette explication a peu de chance d'être fondée, car nous utilisons une architecture *Resnet50* ayant fait ses preuves dans diverses tâches de classification biomédicales avec plus de 25 millions de paramètres, ce qui est suffisant pour détecter une tache centrale sur des images de petite résolution. Un seconde raison expliquant ces erreurs pourrait être la présence de bruit ou d'erreurs dans le jeu de données de base (erreurs d'annotations). Nous avons alors exploré comment nous pouvions, grâce au modèle entrainé, identifier de potentielles erreurs d'annotations dans notre jeu de données.

La figure 8.18 résume la démarche utilisée pour identifier des erreurs dans le jeu de données. À partir de la prédiction de la classe des 16 787 images, nous avons filtré pour ne garder que les images où le label (annotation par l'expert) était discordant avec la prédiction du modèle ET où le modèle avait un fort niveau de confiance dans la prédiction (>85%). Au total, cela représente 228 images, soit 1.66% du jeu de donnée de base.

La figure 8.19 présente huit exemples pris au hasard parmi les 228 images ayant une prédic-

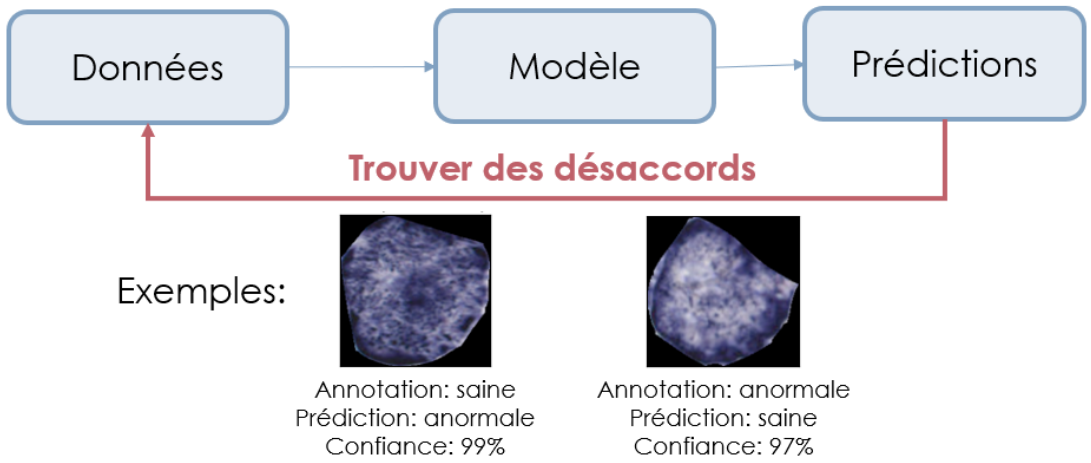


FIGURE 8.18 – **Méthode d'identification des erreurs d'annotation potentielles.** À partir des annotations et des prédictions du modèle sur le jeu de données, nous extrayons les images pour lesquelles le modèle prédit une classe contraire à l'annotation, avec un haut taux de confiance dans la prédiction.

tion discordante avec l'annotation et un haut niveau de confiance du modèle. On observe par exemple pour la fibre n°1 et la fibre n°8, qu'elles ont été annotées par l'expert comme saines et prédites comme anormales par le modèle. En regardant l'image de la fibre, on observe une tache centrale caractéristique de fibres ayant une répartition mitochondriale anormale. Ceci laisse donc penser qu'il y a eu une erreur d'annotation pour ces deux fibres, considérée à tort comme saine. De même, pour la fibre n°2, annotée comme anormale, mais prédite comme saine, on observe la présence d'un marquage homogène sans agglutination caractéristique du marquage au centre, il n'y a donc visiblement pas de raison pour la classer comme anormale. À l'inverse, pour l'image n°5, annotée comme saine, la prédiction du modèle est "anormale" avec un taux de confiance de 96%, malgré une apparence sombre, il ne semble pas y avoir de marquage particulièrement prononcé au centre de la fibre comparé à la périphérie, ce qui correspond donc probablement à une vraie erreur du modèle.

Ces résultats montrent que le modèle est capable d'identifier des erreurs humaines lors de l'évaluation des fibres. Ces erreurs d'annotation peuvent être dues à des inattentions ou des erreurs de clic lors du travail manuel et répétitif d'annotation des 16 787 images par les experts. La détection de potentielles erreurs par le modèle et leur ré-annotation pourrait permettre d'améliorer le jeu de données utilisé pour l'entrainement du modèle et ainsi obtenir de meilleures performances de classification.

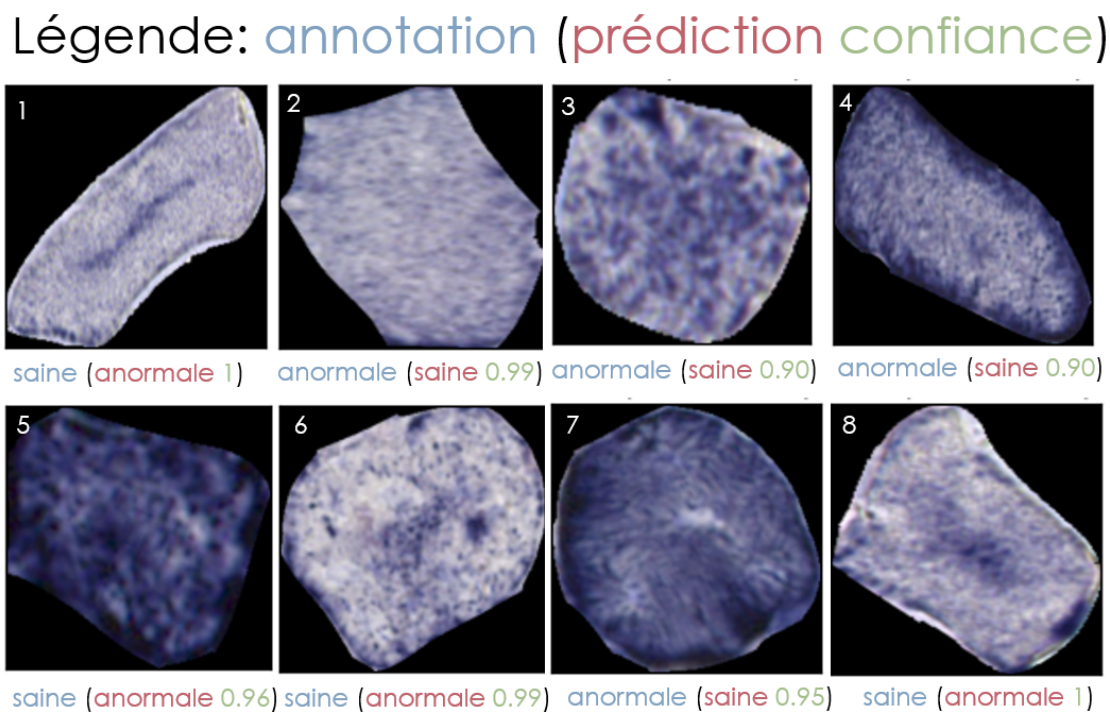


FIGURE 8.19 – **Exemple de fibres ayant un label prédit contraire à l'annotation avec un haut niveau de confiance du modèle**. L'annotation par l'expert est écrite en bleue, la prédiction du modèle en rouge et sa confiance dans la prédiction en vert (comprise entre 0 et 1). Les fibres n°1, 2 et 8 semblent avoir une mauvaise annotation par l'expert. La fibre n°5 semble être bien annotée et est probablement une erreur réelle du modèle.

8.4 Déploiement de la plateforme

Pour faciliter l'utilisation et la diffusion de **MyoQuant**, nous avons développé l'outil sous deux formes : un outil en ligne de commande et une version de démonstration en ligne.

8.4.1 Outil en ligne de commande

Sous sa forme d'outil en ligne de commande, **MyoQuant** est téléchargeable comme bibliothèque Python disponible dans le répertoire officiel PyPI à l'adresse https://pypi.org/project/myoquant/. La version en ligne de commande permet de traiter un grand nombre d'images et/ou des images de coupes complètes trop grandes pour être traitées *via* une interface en ligne. Ceci permet de traiter les images sur des serveurs de calcul équipés de GPU afin d'accélérer le processus d'analyse et de sauvegarder l'ensemble des résultats.

8.4.2 Version de démonstration en ligne

MyoQuant est aussi disponible sous forme d'interface de démonstration en ligne développée grâce à *Streamlit*. Cette interface en ligne est utile pour faire des démonstrations de l'outil de façon visuelle, notamment grâce à la génération des différentes figures présentée ci-dessus améliorant l'explicabilité des classifications (analyse de la centralisation pour la coloration HE, histogrammes et courbe de densité pour la coloration ATPase, méthode Gradcam pour le modèle SDH). Cette interface permet aussi de tester l'outil sur des images de petites tailles afin d'évaluer quels seraient les meilleurs paramètres pour obtenir une classification de qualité. **MyoQuant**-*Streamlit* est disponible en ligne à l'adresse https://lbgi.fr/MyoQuant/ ainsi que sur *HuggingFace Space* https://huggingface.co/spaces/corentinm7/MyoQuant.

8.5 Limites et perspectives de développement

L'outil **MyoQuant** permet de quantifier des marqueurs pathologiques dans trois des cinq colorations réalisées en routine pour le diagnostic des MC. Pour l'instant, **MyoQuant** n'inclus pas de méthode dédiée à l'analyse des coupes à la coloration TG pour détecter les agrégats protéiques ni de méthode pour la coloration NADH pouvant mettre en évidence la présence de *cores*. Des modèles IA suivant la même méthodologie que ce que nous avons développé pour la coloration SDH devront être développés en conséquence. Il est aussi nécessaire pour certaines colorations d'élargir le champ des marqueurs pathologiques quantifiés. Par exemple, pour la coloration HE, la taille des fibres pour la détection de fibres atrophiques est aussi évaluée en routine lors du diagnostic en complément de la position des noyaux.

De plus, **MyoQuant** a majoritairement été développé à partir de données issues de biopsies musculaires de souris. Il serait intéressant de tester **MyoQuant** sur un jeu de données de biopsies humaines quantifiées manuellement par un expert pour évaluer son niveau de performance sur des données humaines. En cas de performance satisfaisante, il serait alors possible d'analyser de façon massive les données de patients afin d'établir de potentiels seuils pour chaque marqueur pathologique pour le diagnostic automatique de MC.

Enfin, sur un plan technique, il serait intéressant de développer une interface en ligne pour **MyoQuant** liée à des GPU pour le traitement de coupes complètes en un temps raisonnable sans avoir à utiliser la ligne de commande, qui est un frein majeur pour un public non expérimenté en programmation, ce qui est typiquement le cas lors du diagnostic.

Quatrième partie

DISCUSSIONS

Chapitre 9

Discussions et ouvertures

Dans les chapitres précédents, nous avons présenté les trois outils (**IMPatienT**, **NLMyo** et **MyoQuant**) que nous avons développés pour exploiter les données multimodales de patients atteints de myopathies congénitales. Bien que ces méthodes soient fonctionnelles, elles présentent des challenges et des perspectives d'amélioration à la fois biologiques et techniques. Au niveau biologique (interprétation des résultats), nous allons discuter de l'intégration des données génomiques comme modalité supplémentaire ainsi que de la question de la mise en relation des différentes modalités entre elles (clinique, histologique et génétique). Au niveau technique, il est important d'aborder la question de l'explicabilité des systèmes IA, des ressources nécessaires à la mise en place de systèmes IA et des aspects législatifs de la mise en place de tels systèmes pour le traitement de données de santé. Enfin, une dernière perspective de valorisation des travaux sera abordée concernant l'intégration des outils dans un projet de création de produit combinant les outils en un seul point d'accès unique.

9.1 Intégration de nouvelles modalités de données : les données génomiques

Les outils développés se sont concentrés majoritairement sur l'exploitation des comptes rendus médicaux et des données de type imagerie. Le traitement des données génomiques dans le système **IMPatienT** est sommaire : il est possible d'associer un gène muté responsable de la maladie (et une mutation) pour un patient et de filtrer les symptômes en fonction du gène muté d'intérêt. Cependant, dans les maladies rares et génétiques, une difficulté majeure est justement de trouver la mutation causant la maladie. Pour rappel, 50% des patients atteints de myopathies congénitales n'ont pas de diagnostic génétique à ce jour. Ainsi il serait intéressant d'ajouter au panel d'outils développés ici des outils capables d'analyser des données de séquençage pour détecter et prioriser des mutations potentiellement responsables de maladies génétiques.

Par exemple, la figure 9.1 présente une façon d'intégrer les données génomiques. L'utilisateur pourrait joindre au dossier du patient un fichier VCF listant les variants trouvés dans son génome. À partir de cette liste, il serait possible de filtrer les variants pour ne garder que les variants liés aux gènes des myopathies congénitales puis de classer les variants par niveau de pathogénicité (en intégrant des outils de prédiction de pathogénicité comme MISTIC (Chennen et al., 2020)). Enfin, il serait possible de prioriser les variants de cette liste en croisant la liste avec les données phénotypiques et histologiques. Par exemple dans le cadre des myopathies à

némaline, la présence de bâtonnets dans les fibres musculaires est souvent liée à une mutation dans les gènes NEB, TPM2, TPM3 ou ACTA1. Ainsi on pourrait filtrer la liste de variants pour ne conserver que les variants pathogènes dans ces gènes. Cette approche pourrait permettre de filtrer rapidement de potentiels variants responsables de la maladie génétique du patient.

Cependant, la recherche exhaustive de variants génétiques par séquençage complet du génome n'est pas universellement disponible en France et donc les données sont limitées. Des initiatives comme le Plan France Médecine Génomique 2025 visent à démocratiser la médecine génomique pour le diagnostic des patients.

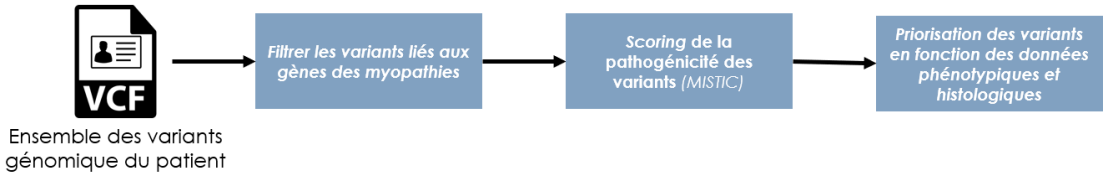


FIGURE 9.1 – **Exemple de méthode d'intégration des données génomiques.** Une méthode d'intégration des données génomiques à **IMPatienT** serait de permettre le dépôt d'un fichier VCF listant l'ensemble des variants dans le génome du patient. Ensuite **IMPatienT** filtrerait les variants pour ne sélectionner que les variants liés à des gènes de MC et ayant un score de pathogénicité élevé. Enfin une dernière étape de filtre consisterait à utiliser les observations phénotypiques et histologiques pour accorder plus d'importance aux variants dans les gènes causant ces phénotypes.

9.2 Mise en relation des modalités

Jusqu'ici, nous avons traité les différentes modalités des données dans les maladies rares (données cliniques, histologiques et génétiques) séparément, par une approche en silo. Nous avons utilisé cette approche en raison de la nature parcellaire et fragmentée des données que nous avions à disposition et du mélange de données humaines (rapport de biopsie) et données de souris (coupe histologique de modèle de MC). Cette approche, bien qu'utile, limite nos capacités à réaliser des découvertes transversales à partir de données multimodales. En effet, chaque type de données ne révèle qu'une facette de la maladie.

Il est probable que ce n'est que par le croisement des différentes modalités que l'on puisse accéder à une meilleure compréhension des myopathies congénitales. Par exemple, une caractéristique clinique pourrait être observée plus fréquemment chez des patients avec un profil histologique et génétique particulier. La détection de telles interactions pourrait permettre un diagnostic facilité et plus rapide sur la simple base des symptômes cliniques.

Pour cela, il est nécessaire de constituer un jeu de données à haute valeur ajoutée. Ce jeu de données comprendrait une cohorte de patients pour lesquels les informations cliniques, histologiques et génétiques complètes aient été récupérées. L'analyse de ce jeu de données avec les outils développés ici pourrait permettre de découvrir des corrélations de phénotypes entre les modalités, d'établir des seuils pour les marqueurs pathologiques quantifiés et de mettre en évidence des critères discriminants pour le diagnostic des myopathies congénitales.

9.3 Explicabilité et exploration des données patient

L'explicabilité des modèles d'IA est un enjeu majeur, notamment en médecine quand ces modèles sont employés pour le diagnostic. Parmi les méthodes que nous avons développées, le niveau d'explicabilité est variable.

La plateforme d'annotation et d'exploration **IMPatienT**, qui est relativement couteuse en travail manuel de mise en forme des données, offre l'avantage de permettre une classification par IA transparente et explicable. Comme vu dans les chapitres 5 et 6, l'importance de chaque symptôme dans le cadre d'un modèle de classification peut être évaluée, ce qui permet une confiance accrue dans les prédictions.

MyoQuant utilise des réseaux de neurones profonds pour la quantification, ces réseaux sont en général considérés comme des boites noires. Cependant ici, le réseau de neurones n'est pas utilisé pour faire du diagnostic, mais pour extraire et quantifier des marqueurs pathologiques sur lesquels les prédictions peuvent être réalisées. Ainsi on a un système explicable, car le système en boite noire n'est utilisé que pour quantifier des symptômes, ce qui est vérifiable à tout moment en comptant manuellement.

Cependant, concernant **NLMyo**, qui repose sur les LLMs, l'explicabilité est moindre. En effet, le système de classification de **NLMyo** repose sur les modèles LLMs d'*embedding* qui sont des boites noires totales, la signification des valeurs du vecteur issues des modèles d'*embedding* est inconnue. Le système permet donc une classification automatique et instantanée, mais il n'est pas transparent ni explicable : il est impossible de savoir sur quels critères la classification a été réalisée.

Les recherches en termes d'explicabilité doivent se poursuivre, notamment pour envisager l'utilisation de modèles IA pour réaliser de l'aide au diagnostic. Dans cette thèse, nous avons exploré trois niveaux d'explicabilité via les différentes méthodes développées. Une solution mise en place ici pour utiliser les modèles de réseau de neurones, considérés comme des boites noires, a été de les utiliser pour faire simplement de l'extraction et quantification de marqueurs à partir d'image avec **MyoQuant** et de l'annotation structurée de comptes rendus de patients grâce aux modèles de langage pour le données textuelles avec **NLMyo**. Et, enfin, utiliser ces données extraites par des algorithmes explicables (xIA) de classification pour le diagnostic et l'extraction de connaissance.

9.4 Ressources informatiques et déploiement de méthodes IA

Les approches IA révolutionnent la façon d'exploiter les données biomédicales. Cependant, elles représentent un cout en calcul important qui met en évidence un challenge technique pour la mise en place de telles approches.

En termes d'analyse d'images, notamment pour les images histologiques, la taille des images à analyser est particulièrement importante. Ainsi sur une machine classique, avec un processeur (CPU) et sans GPU, l'analyse risque d'être bien trop lente. De plus, la taille des images histologiques peut souvent être importante au point de dépasser la mémoire disponible sur les GPU classiques. Cette grande taille des images histologiques induit la nécessité de matériels ou de méthodes spécifiques, capables de gérer des volumes de données importants.

Pour l'analyse de texte, les LLMs comme leur nom l'indique sont des modèles très grands et lourds à héberger (plus de 170 milliards de paramètres pour *GPT-3*). Ils requièrent donc du matériel spécialisé et couteux, c'est-à-dire plusieurs GPU haut de gamme. Cependant, les efforts d'optimisation en cours comme la *quantisation* sur 4 bits ont permis de réduire les ressources nécessaires pour héberger de tels modèles. En seulement 2 mois après la sortie de

grands modèles *open-source* comme *LLaMA*, il est maintenant possible d'héberger ces modèles de langages sur des machines avec seulement 8 GO de mémoire vive. Ces modèles plus petits et optimisés ont cependant un coût : une qualité de résultats réduite et un temps de génération plus long, les rendant difficilement utilisables en routine. Cependant, nous espérons que ces progrès continueront avec un rythme soutenu, rendant ces modèles de plus en plus accessibles.

En règle générale, les modèles IA restent difficiles à distribuer et à déployer. L'écosystème informatique nécessaire pour héberger et faire tourner des modèles de réseaux de neurones profonds en production reste complexe et peut être difficile à gérer et à maintenir. Plutôt qu'un système distribué où chaque client pourrait faire tourner ses modèles, un système centralisé qui mutualise une installation des modèles et les ressources nécessaires à leur fonctionnement serait bénéfique et plus viable à long terme en termes de couts, stabilité et maintenance.

9.5 RGPD et traitement de données de santé

La solution aux difficultés de l'hébergement des modèles IA peut être l'utilisation d'hébergeurs externes. Par exemple, pour les LLMs, des fournisseurs de modèles et d'hébergement comme *OpenAI* via leur API permettent d'utiliser ces modèles sans avoir besoin de ressources de calcul, par facturation à la requête. Toutefois, cette solution présente un challenge majeur : la confidentialité des données.

En particulier dans le domaine de la santé, les données font l'objet d'un traitement spécifique à cause de leur sensibilité. Leur hébergement et traitement demande un haut niveau de confidentialité et de protection. En particulier en Europe, le RGPD qui assure la protection de données personnelles, requiert des exigences strictes en termes de sécurité pour le traitement des données de santé. En France, pour utiliser un service informatique externe pour traiter ou stocker des données, celui-ci doit être certifié HDS (hébergeur de données de santé). Ces règles ne concernent pas que les données directement identifiantes comme les noms et dates de naissance, mais aussi des données considérées comme potentiellement identifiantes : résultats génétiques, images histologiques, profil phénotypique.

La nécessité de l'application du RGPD et la certification HDS représentent un frein dans la mise en place et l'hébergement de modèles IA innovant pour la recherche biomédicale auprès des équipes de recherche. Ainsi, dans le cadre d'un développement de logiciel pour traiter les données de santé (comme **IMPatienT**, **MyoQuant** et **NLMyo**), il y a deux options possibles. La première est l'utilisation de services externes (pour l'hébergement de données ou le calcul de modèles IA). Dans ce cas, nous sommes restreints à l'utilisation de services hébergés en Europe et certifiés HDS, ce qui n'est pas toujours possible. Par exemple, dans le cas du modèle LLMs GPT-3.5 de OpenAI, il n'y a pas à ce jour d'hébergement publiquement accessible en Europe certifié HDS. *Azure OpenAI Service* est un hébergement européen du modèle OpenAI GPT-3.5, certifié HDS, mais il n'est disponible que sur invitation pour le moment. La deuxième voie consiste à héberger localement les modèles et le stockage des données, ce qui peut se révéler difficile sur un plan technique, plus couteux et moins performant.

Malgré l'ensemble de ces défis, il est donc important de penser une organisation des outils IA permettant d'être conforme aux législations pour permettre de tirer parti des bénéfices de l'IA appliquée au domaine biomédical.

9.6 Valorisation des travaux : intégration des outils en un point d'accès unique

Dans l'optique de valoriser les travaux présentés dans cette thèse et de démocratiser l'utilisation de l'IA pour l'analyse des données biomédicales, il est nécessaire d'intégrer l'ensemble des outils développés en un point d'accès unique. Cette vision a été encouragée par la *SATT Connectus* à travers le challenge *Mature your PhD* visant à transformer les travaux de recherches en produits innovants.

L'objectif principal est de concilier les trois outils **IMPatienT**, **NLMyo** et **MyoQuant** en une plateforme unique pour faciliter leur utilisation par les chercheurs. Ainsi l'application web et la base de données **IMPatienT** pourraient servir de pilier de base avec l'intégration de **NLMyo** et **MyoQuant** pour accélérer l'analyse des données.

Ce projet mets en évidence des défis majeurs déjà identifiés dans les sections ci-dessus, notamment concernant les besoins en ressources de calculs des modèles IA, la complexité technique de leur maintenance et les questions de confidentialité des données au regard du RGPD. Cependant, pour relever ces défis, il serait possible de structurer cette plateforme en deux parties : un client et un point central de calcul.

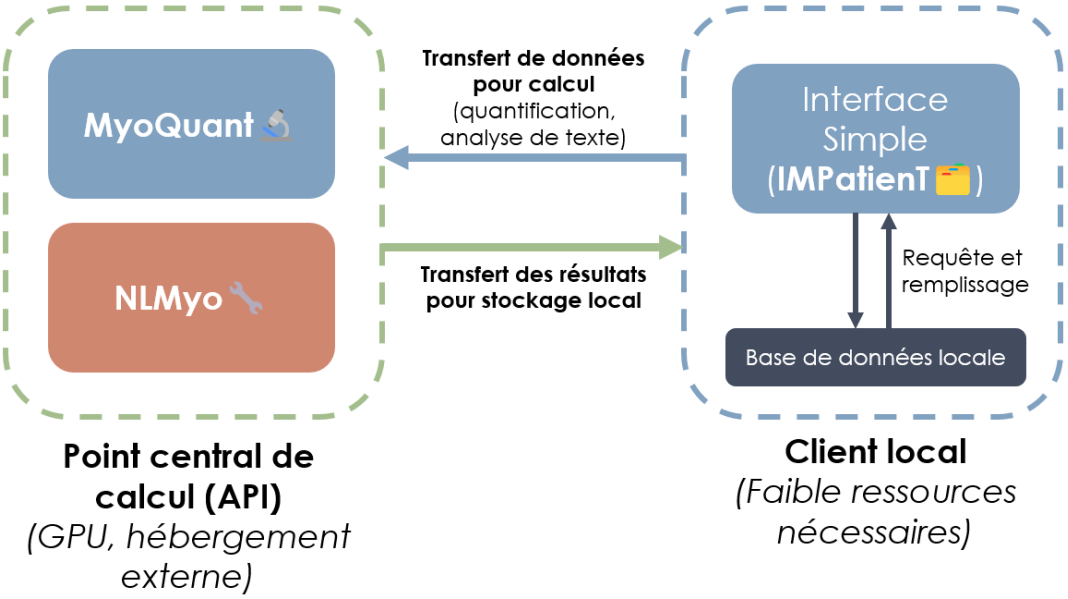


FIGURE 9.2 – **Architecture de la plateforme unique en deux parties.** Un client local (à droite) servant d'interface et hébergeant les données et un serveur de calcul distant mutualisé (à gauche) mettant à disposition les modèles IA et les ressources de calcul nécessaires.

La figure 9.2 présente l'architecture prévue pour la plateforme unifiée. Le client serait une interface simple permettant d'interagir avec une base de données hébergée localement chez l'utilisateur afin de maintenir la confidentialité et le contrôle des données. À partir de cette interface, il serait possible d'ajouter des patients et de lancer des analyses des données enregistrées (quantification d'images par **MyoQuant**, ou analyse de texte par **NLMyo**). Ces analyses de données seraient alors déléguées au point central de calcul. Ainsi le client transfèrerait uniquement les données nécessaires à l'IA au point central de calcul, qui, quant à lui, fournirait les résultats à

stocker localement au client.

Cette architecture est adaptée, car elle permet à la fois de garantir la confidentialité des données, donnant le contrôle total de leurs données aux chercheurs. Mais aussi elle permet de mutualiser les couts en ressources informatiques des modèles IA avec un point central pour plusieurs clients, dont le déploiement et la maintenance seraient assurés par notre expertise. Au final, cette architecture adaptée en deux parties, qui intègre les trois outils présentés dans cette thèse, permet à la fois de répondre aux challenges techniques tout en facilitant l'utilisation d'IA au service de l'analyse des données biomédicales pour les chercheurs.

Bibliographie

Abdallah, Y. M. Y. (2017). History of medical imaging. *Archives of Medicine and Health Sciences*, *5*(2), 275. https://doi.org/10.4103/amhs.amhs_97_17

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, *28*(9), 1773-1784. https://doi.org/10.1038/s41591-022-01981-2

Adrien Treuille, Amanda Kelly & Thiago Teixeira. (2018). *Streamlit : A faster way to build and share data apps.* Récupérée 9 juin 2023, à partir de https://github.com/streamlit/streamlit

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna : A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alan Pestronk. (2022, octobre 26). *NEUROMUSCULAR DISEASE CENTER*. Récupérée 26 juin 2023, à partir de https://neuromuscular.wustl.edu/index.html

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W. J., Neves, C., Patterson, A., Porter, M., Tarkowska, A., Loranger, B., Avondo, J., Lagerstedt, I., Lianas, L., Leo, S., Hands, K., Hay, R. T., … Swedlow, J. R. (2012). OMERO : flexible, model-driven data management for experimental biology. *Nature Methods*, *9*(3), 245-253. https://doi.org/10.1038/nmeth.1896

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org : Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*, D789-D798. https://doi.org/10.1093/nar/gku1205

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI) : concepts, taxonomies, opportunities and challenges toward responsible AI. *arXiv :1910.10045 [cs]*. Récupérée 15 juin 2021, à partir de http://arxiv.org/abs/1910.10045

Aurélien Géron. (2019, septembre). *Hands-on machine learning with scikit-learn, keras, and TensorFlow, 2nd edition* (O'Reilly). Récupérée 21 juin 2023, à partir de https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, *18*(10), 1196-1203. https://doi.org/10.1038/s41592-021-01252-x

Ben Goldacre. (2022, avril 21). *Better, broader, safer : using health data for research and analysis* [GOV.UK]. Récupérée 16 juin 2023, à partir de https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis

Benarroch, L., Bonne, G., Rivier, F., & Hamroun, D. (2023). The 2023 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscular disorders : NMD*, *33*(1), 76-117. https://doi.org/10.1016/j.nmd.2022.12.002

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, juillet 22). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023, avril 13). Sparks of Artificial General Intelligence : Early experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712

Burr, D. B., & Allen, M. R. (Éd.). (2019, janvier 1). *Basic and applied bone biology* (2nd Edition). Academic Press. https://doi.org/10.1016/B978-0-12-813259-3.01001-0

Cadot, B., & Romero, N. B. (2022). *Atlas du muscle* [Institut de Myologie]. Récupérée 25 avril 2022, à partir de https://www.institut-myologie.org/recherche/centre-dexploration-et-devaluation-neuromusculaire/laboratoire-dhistopathologie-dr-n-b-romero/atlas-du-muscle/

Cassandrini, D., Trovato, R., Rubegni, A., Lenzi, S., Fiorillo, C., Baldacci, J., Minetti, C., Astrea, G., Bruno, C., Santorelli, F. M., & Italian Network on Congenital Myopathies. (2017). Congenital myopathies : clinical phenotypes and new diagnostic tools. *Italian Journal of Pediatrics*, *43*(1), 101. https://doi.org/10.1186/s13052-017-0419-z

Chase Harrison. (2022, octobre 17). *LangChain : building applications with LLMs through composability* [GitHub]. Récupérée 9 juin 2023, à partir de https://github.com/hwchase17/langchain

Chen, R., & Snyder, M. (2013). Promise of Personalized Omics to Precision Medicine. *Wiley interdisciplinary reviews. Systems biology and medicine*, *5*(1), 73-82. https://doi.org/10.1002/wsbm.1198

Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chennen, K., Weber, T., Lornage, X., Kress, A., Böhm, J., Thompson, J., Laporte, J., & Poch, O. (2020). MISTIC : A prediction tool to reveal disease-relevant deleterious missense variants. *PloS One*, *15*(7), e0236962. https://doi.org/10.1371/journal.pone.0236962

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., & Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research*, *22*(3), 490-503. https://doi.org/10.1038/cr.2012.15

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023, mars). Vicuna : An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chollet, F., et al. (2015). Keras. https://keras.io

Chollet, F. (2023, juin 21). *Keras documentation : keras applications*. Récupérée 21 juin 2023, à partir de https://keras.io/api/applications/

Claeys, K. G. (2020). Congenital myopathies : an update. *Developmental Medicine and Child Neurology*, *62*(3), 297-302. https://doi.org/10.1111/dmcn.14365

Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2017, avril 7). Hierarchical Clustering : Objective Functions and Algorithms. https://doi.org/10.48550/arXiv.1704.02147

Dai, X., & Shen, L. (2022). Advances and Trends in Omics Technology Development. *Frontiers in Medicine*, *9*. Récupérée 19 juillet 2023, à partir de https://www.frontiersin.org/articles/10.3389/fmed.2022.911861

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*(3), 122-135. https://doi.org/10.1108/LR-06-2015-0061

de Mello, B. H., Rigo, S. J., da Costa, C. A., da Rosa Righi, R., Donida, B., Bez, M. R., & Schunke, L. C. (2022). Semantic interoperability in health records standards : a systematic literature review. *Health and Technology*, *12*(2), 255-272. https://doi.org/10.1007/s12553-022-00639-w

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*, 248-255.

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022, novembre 10). LLM.int8() : 8-bit Matrix Multiplication for Transformers at Scale. https://doi.org/10.48550/arXiv.2208.07339

Dettmers, T., & Zettlemoyer, L. (2023, février 27). The case for 4-bit precision : k-bit Inference Scaling Laws. Récupérée 24 juin 2023, à partir de http://arxiv.org/abs/2212.09720

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226-231.

European Organization For Nuclear Research & OpenAIRE. (2013). Zenodo. https://doi.org/10.25495/7GXK-RD71

Frauke Rudolf. (2023). AI software at least as good as radiologists at detecting TB from chest x-rays. Récupérée 19 juin 2023, à partir de https://www.eurekalert.org/news-releases/986259

Fukushima, K. (1980). Neocognitron : a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193-202. https://doi.org/10.1007/BF00344251

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022, novembre 22). Red Teaming Language Models to Reduce Harms : Methods, Scaling Behaviors, and Lessons Learned. https://doi.org/10.48550/arXiv.2209.07858

García, C. G., & Álvarez-Fernández, E. (2022). What Is (Not) Big Data Based on Its 7Vs Challenges : A Survey. *Big Data Cogn. Comput.*, *6*(4), 158. https://doi.org/10.3390/bdcc6040158

Gineste, C., & Laporte, J. (2023). Therapeutic approaches in different congenital myopathies. *Current Opinion in Pharmacology*, *68*, 102328. https://doi.org/10.1016/j.coph.2022.102328

Gómez Oca, R. (2021, décembre 7). *Physiological role of muscle-specific dynamin-2 isoform and its impact on centronuclear myopathy pathology and treatment* (These de doctorat). Strasbourg. Récupérée 27 juin 2023, à partir de https://www.theses.fr/2021STRAJ065

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., & Bojanowski, P. (2021, mars 5). Self-supervised Pretraining of Visual Features in the Wild. Récupérée 23 juin 2023, à partir de http://arxiv.org/abs/2103.01988

Graber, M. L., Byrne, C., & Johnston, D. (2017). The impact of electronic health records on diagnosis. *Diagnosis (Berlin, Germany)*, *4*(4), 211-223. https://doi.org/10.1515/dx-2017-0012

Guan, X., Goddard, M. A., Mack, D. L., & Childers, M. K. (2016). Gene therapy in monogenic congenital myopathies. *Methods (San Diego, Calif.)*, *99*, 91-98. https://doi.org/10.1016/j.ymeth.2015.10.004

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., & Song, D. (2023, mai 25). The False Promise of Imitating Proprietary LLMs. https://doi.org/10.48550/arXiv.2305.15717

Hanauer, D. A., Mei, Q., Law, J., Khanna, R., & Zheng, K. (2015). Supporting information retrieval from electronic health records : A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *Journal of Biomedical Informatics*, *55*, 290-300. https://doi.org/10.1016/j.jbi.2015.05.003

He, K., Zhang, X., Ren, S., & Sun, J. (2015, décembre 10). Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135-175. Récupérée 15 octobre 2020, à partir de https://www.jstor.org/stable/185169

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding.

Hipp, R. D. (2020, août). *SQLite* (Version 3.31.1). https://www.sqlite.org/index.html

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, *9*, 1735-80. https://doi.org/10.1162/neco.1997.9.8.1735

Hölscher, D. L., Bouteldja, N., Joodaki, M., Russo, M. L., Lan, Y.-C., Sadr, A. V., Cheng, M., Tesar, V., Stillfried, S. V., Klinkhammer, B. M., Barratt, J., Floege, J., Roberts, I. S. D., Coppo, R., Costa, I. G., Bülow, R. D., & Boor, P. (2023). Next-generation morphometry for pathomics-data mining in histopathology. *Nature Communications*, *14*(1), 470. https://doi.org/10.1038/s41467-023-36173-0

Huang, K., Bi, F.-F., & Yang, H. (2021). A Systematic Review and Meta-Analysis of the Prevalence of Congenital Myopathy. *Frontiers in Neurology*, *12*, 761636. https://doi.org/10.3389/fneur.2021.761636

Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology*, *147*(2), 226-238.2. Récupérée 21 juin 2023, à partir de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1357023/

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574-591. Récupérée 21 juin 2023, à partir de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/

Ishwarappa & Anuradha, J. (2015). A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, *48*, 319-324. https://doi.org/10.1016/j.procs.2015.04.188

Ismail, L., Materwala, H., Karduck, A. P., & Adem, A. (2020). Requirements of Health Data Management Systems for Biomedical Care and Research : Scoping Review. *Journal of Medical Internet Research*, *22*(7), e17508. https://doi.org/10.2196/17508

Janssen, I., Heymsfield, S. B., Wang, Z. M., & Ross, R. (2000). Skeletal muscle mass and distribution in 468 men and women aged 18-88 yr. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *89*(1), 81-88. https://doi.org/10.1152/jappl.2000.89.1.81

Jumper, J. (2021a). Applying and improving AlphaFold at CASP14. *Proteins : Structure, Function, and Bioinformatics*, *89*(12), 1711-1721. https://doi.org/10.1002/prot.26257

Jumper, J. (2021b). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 1-11. https://doi.org/10.1038/s41586-021-03819-2

Jungbluth, H., Treves, S., Zorzato, F., Sarkozy, A., Ochala, J., Sewry, C., Phadke, R., Gautel, M., & Muntoni, F. (2018). Congenital myopathies : disorders of excitation-contraction coupling and muscle contraction. *Nature Reviews. Neurology*, *14*(3), 151-167. https://doi.org/10.1038/nrneurol.2017.191

Karollus, A., Mauermeier, T., & Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, *24*(1), 56. https://doi.org/10.1186/s13059-023-02899-9

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM : A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. Récupérée 19 juillet 2023, à partir de https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Keith J. Dreyer, Christoph Wald, Bibb Allen Jr, Bernardo C. Bizzo, Sheela Agarwal, Judy W. Gichoya & Jay Patti. (2023, juin 19). *ACR Data Science Institute AI Central*. Récupérée 19 juin 2023, à partir de https://aicentral.acrdsi.org/

Ker, J., Bai, Y., Lee, H. Y., Rao, J., & Wang, L. (2019). Automated brain histology classification using machine learning. *Journal of Clinical Neuroscience*, *66*, 239-245. https://doi.org/10.1016/j.jocn.2019.05.019

Kim, B.-H., Deng, Z., Yu, P. S., & Ganapathi, V. (2022, octobre 28). Can Current Explainability Help Provide References in Clinical Notes to Support Humans Annotate Medical Codes ? Récupérée 23 juin 2023, à partir de http://arxiv.org/abs/2210.15882

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, *49*, D1207-D1217. https://doi.org/10.1093/nar/gkaa1043

Lace, B., Micule, I., Kenina, V., Setlere, S., Strautmanis, J., Kazaine, I., Taurina, G., Murmane, D., Grinfelde, I., Kornejeva, L., Krumina, Z., Sterna, O., Radovica-Spalvina, I., Vasiljeva, I., Gailite, L., Stavusis, J., Livcane, D., Kidere, D., Malniece, I., & Inashkina, I. (2022). Overview of neuromuscular disorder molecular diagnostic experience for the population of latvia. *Neurology Genetics*, *8*(3). https://doi.org/10.1212/NXG.0000000000000685

Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., Hernandez-Ferrer, C., Luengo, C., Martínez, I., Papakonstantinou, A., Picó-Amador, D., Protasio, J., Thompson, R., Tonda, R., Bayés, M., Bullich, G., Camps-Puchadas, J., Paramonov, I., Trotta, J.-R., … Beltran, S. (2022). The RD-Connect Genome-Phenome Analysis Platform : Accelerating diagnosis, research, and gene discovery for rare diseases. *Human Mutation*, *43*(6), 717-733. https://doi.org/10.1002/humu.24353

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Ha, P. (1998). Gradient-based learning applied to document recognition.

LeCun, Y., & Misra, I. (2021, avril 3). *Self-supervised learning : the dark matter of intelligence*. Récupérée 5 janvier 2022, à partir de https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Leijnen, F. v. V., Stefan. (2016, septembre 14). *The neural network zoo* [The asimov institute]. Récupérée 20 juin 2023, à partir de https://www.asimovinstitute.org/neural-network-zoo/

Li, T., Wang, Z., Lu, W., Zhang, Q., & Li, D. (2022). Electronic health records based reinforcement learning for treatment optimizing. *Information Systems*, *104*, 101878. https://doi.org/10.1016/j.is.2021.101878

Lianmin Zheng, Ying Sheng, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez & Ion Stoica. (2023, mars 5). *Chatbot arena : benchmarking LLMs in the wild with elo ratings | LMSYS org*. Récupérée 9 juin 2023, à partir de https://lmsys.org/blog/2023-05-03-arena

Lornage, X. (2019, septembre 13). *Identification and functional characterization of novel genes implicated in congenital myopathies* (These de doctorat). Strasbourg. Récupérée 27 juin 2023, à partir de https://www.theses.fr/2019STRAJ067

Luis Serrano. (2023, janvier 18). *What are word and sentence embeddings ?* [Context by cohere]. Récupérée 23 juin 2023, à partir de https://txt.cohere.com/sentence-word-embeddings/

Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579-2605. Récupérée 19 juillet 2023, à partir de http://jmlr.org/papers/v9/vandermaaten08a.html

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R

MacQueen, J. (1967, janvier 1). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics* (p. 281-298). University of California Press. Récupérée 19 juillet 2023, à partir de https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Maiella, S., Rath, A., Angin, C., Mousson, F., & Kremp, O. (2013). Orphanet et son réseau : où trouver une information validée sur les maladies rares. *Revue Neurologique*, *169*, S3-S8. https://doi.org/10.1016/S0035-3787(13)70052-3

Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.-M., Kainz, P., Geurts, P., & Wehenkel, L. (2016). Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics*, *32*(9), 1395-1401. https://doi.org/10.1093/bioinformatics/btw013

Mark Otto & Jacob Thornton. (2011, août 19). *Bootstrap : The most popular HTML, CSS, and JavaScript framework for developing responsive, mobile first projects on the web.* (Version 5). Récupérée 9 juin 2023, à partir de https://github.com/twbs/bootstrap

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... Xiaoqiang Zheng. (2015). TensorFlow : Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

Martínez-García, M., & Hernández-Lemus, E. (2022). Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*, *8*, 784455. https://doi.org/10.3389/fmed.2021.784455

McInnes, L., Healy, J., & Melville, J. (2020, septembre 17). UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426

Momeni, Z., Hassanzadeh, E., Saniee Abadeh, M., & Bellazzi, R. (2020). A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, *107*, 103466. https://doi.org/10.1016/j.jbi.2020.103466

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB : Massive Text Embedding Benchmark. *arXiv preprint arXiv :2210.07316*. https://doi.org/10.48550/ARXIV.2210.07316

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy : Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319-327. https://doi.org/10.18653/v1/W19-5034

North, K. N., Wang, C. H., Clarke, N., Jungbluth, H., Vainzof, M., Dowling, J. J., Amburgey, K., Quijano-Roy, S., Beggs, A. H., Sewry, C., Laing, N. G., Bönnemann, C. G., & International Standard of Care Committee for Congenital Myopathies. (2014). Approach to the diagnosis of congenital myopathies. *Neuromuscular disorders : NMD*, *24*(2), 97-116. https://doi.org/10.1016/j.nmd.2013.11.003

Nunn, J. S., Tiller, J., Fransquet, P., & Lacaze, P. (2019). Public Involvement in Global Genomics Research : A Scoping Review. *Frontiers in Public Health*, *7*. Récupérée 17 juin 2023, à partir de https://www.frontiersin.org/articles/10.3389/fpubh.2019.00079

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, mars 4). Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. Récupérée 25 janvier 2022, à partir de http://jmlr.org/papers/v12/pedregosa11a.html

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023, avril 6). *Instruction tuning with GPT-4* [arXiv.org]. Récupérée 24 juin 2023, à partir de https://arxiv.org/abs/2304.03277v1

Prakash, K., Diederich, B., Heintzmann, R., & Schermelleh, L. (2022). Super-resolution microscopy : a brief history and new avenues. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, *380*(2220), 20210110. https://doi.org/10.1098/rsta.2021.0110

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019, janvier 20). CatBoost : unbiased boosting with categorical features. https://doi.org/10.48550/arXiv.1706.09516

Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., & Inoue, I. (2012). Next-generation sequencing : impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, *57*(10), 621-632. https://doi.org/10.1038/jhg.2012.91

Ramírez, S. (2019, décembre). *Typer, build great CLIs. Easy to code. Based on Python type hints.* Récupérée 10 juin 2023, à partir de https://github.com/tiangolo/typer

Raufaste-Cazavieille, V., Santiago, R., & Droit, A. (2022). Multi-omics analysis : Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology. *Frontiers in Molecular Biosciences*, *9*. Récupérée 19 juillet 2023, à partir de https://www.frontiersin.org/articles/10.3389/fmolb.2022.962743

Ray, S., Ahmad, A., Rika, A., Nicholas, B., Jeff, B., Samuel, C., Phil, C., Simon, C., David, E., Sheelagh, H., Dan, J., Rajesh, K., Thomas, K., Dar-Shyang, L., Zongyi, (Robert, M., Chris, N., Michael, R., Marius, R., … Oscar, Z. (2015, juillet). *Tesseract : Open Source OCR Engine*. tesseract-ocr. Récupérée 7 juin 2023, à partir de https://github.com/tesseract-ocr/tesseract

Resig, J. (2006, janvier). *jQuery — New Wave JavaScript*. jQuery. Récupérée 9 juin 2023, à partir de https://github.com/jquery/jquery

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology : a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, *83*(5), 610-615. https://doi.org/10.1016/j.ajhg.2008.09.017

Ronacher, A. (2010, avril 1). *Flask : The Python micro framework for building web applications*. Pallets. Récupérée 9 juin 2023, à partir de https://github.com/pallets/flask

Ronneberger, O., Fischer, P., & Brox, T. (2015, mai 18). U-Net : Convolutional Networks for Biomedical Image Segmentation. https://doi.org/10.48550/arXiv.1505.04597

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386-408. https://doi.org/10.1037/h0042519

Saha, S., Majumdar, D., & Mitra, M. (2022, décembre 14). Explainability of Text Processing and Retrieval Methods : A Critical Survey. Récupérée 23 juin 2023, à partir de http://arxiv.org/abs/2212.07126

Savova, G. K., Tseytlin, E., Finan, S., Castine, M., Miller, T., Medvedeva, O., Harris, D., Hochheiser, H., Lin, C., Chavan, G., & Jacobson, R. S. (2017). DeepPhe : A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. *Cancer Research*, *77*(21), e115-e118. https://doi.org/10.1158/0008-5472.CAN-17-0615

Scalzitti, N. (2021, septembre 29). *Nouvelle stratégie d'annotation des génomes par l'utilisation d'algorithmes d'intelligence artificielle* (These de doctorat). Strasbourg. Récupérée 20 juin 2023, à partir de https://www.theses.fr/2021STRAJ040

Scalzitti, N., Kress, A., Orhand, R., Weber, T., Moulinier, L., Jeannin-Girardon, A., Collet, P., Poch, O., & Thompson, J. D. (2021). Spliceator : multi-species splice site prediction using convolutional neural networks. *BMC Bioinformatics*, *22*(1), 561. https://doi.org/10.1186/s12859-021-04471-3

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM : Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336-359. https://doi.org/10.1007/s11263-019-01228-7

Sheppard, C. J. R. (2021). Structured illumination microscopy and image scanning microscopy : a review and comparison of imaging properties. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, *379*(2199), 20200154. https://doi.org/10.1098/rsta.2020.0154

Sonawane, A. R., Weiss, S. T., Glass, K., & Sharma, A. (2019). Network Medicine in the Age of Biomedical Big Data. *Frontiers in Genetics*, *10*. Récupérée 16 juin 2023, à partir de https://www.frontiersin.org/articles/10.3389/fgene.2019.00294

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008-3021. Récupérée 23 juin 2023, à partir de https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose : a generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100-106. https://doi.org/10.1038/s41592-020-01018-x

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., & Yu, T. (2023, mai 30). One Embedder, Any Task : Instruction-Finetuned Text Embeddings. https://doi.org/10.48550/arXiv.2212.09741

Tajsharghi, H., & Oldfors, A. (2013). Myosinopathies : pathology and mechanisms. *Acta Neuropathologica*, *125*(1), 3-18. https://doi.org/10.1007/s00401-012-1024-2

Tan, M., & Le, Q. V. (2020, septembre 11). EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. https://doi.org/10.48550/arXiv.1905.11946

Tchito Tchapga, C., Mih, T. A., Tchagna Kouanou, A., Fozin Fonzin, T., Kuetche Fogang, P., Mezatio, B. A., & Tchiotsop, D. (2021). Biomedical Image Classification in a Big Data Architecture Using Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2021*, 9998819. https://doi.org/10.1155/2021/9998819

The Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1), iyad031. https://doi.org/10.1093/genetics/iyad031

Toga, A. W., Foster, I. T., Kesselman, C., Madduri, R. K., Chard, K., Deutsch, E. W., Price, N. D., Glusman, G., Heavner, B. D., Dinov, I. D., Ames, J., Horn, J. D. V., Kramer, R., & Hood, L. E. (2015). Big biomedical data as the key resource for discovery science. *J. Am. Medical Informatics Assoc.*, *22*(6), 1126-1131. https://doi.org/10.1093/jamia/ocv077

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023, février 27). *LLaMA : open and efficient foundation language models* [arXiv.org]. Récupérée 21 juin 2023, à partir de https://arxiv.org/abs/2302.13971v1

Urbanowicz, R., Zhang, R., Cui, Y., & Suri, P. (2023). STREAMLINE : a simple, transparent, end-to-end automated machine learning pipeline facilitating data analysis and algorithm comparison. In L. Trujillo, S. M. Winkler, S. Silva & W. Banzhaf (Éd.), *Genetic programming theory and practice XIX* (p. 201-231). Springer Nature. https://doi.org/10.1007/978-981-19-8460-0_9

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection : introduction and review. *Journal of Biomedical Informatics*, *85*, 189-203. https://doi.org/10.1016/j.jbi.2018.07.014

Urbanowicz, R. J., & Moore, J. H. (2015). ExSTraCS 2.0 : Description and Evaluation of a Scalable Learning Classifier System. *Evolutionary intelligence*, *8*(2), 89-116. https://doi.org/10.1007/s12065-015-0128-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, décembre 5). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Victor Dubowitz, Caroline A. Sewry & Anders Oldfors. (2020). *Muscle biopsy. a practical approach 5th edition* (Elsevier). Récupérée 26 juin 2023, à partir de https://www.decitre.fr/livres/muscle-biopsy-9780702074714.html

Walt, S. v. d., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image : image processing in python. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Wang, X., Williams, C., Liu, Z. H., & Croghan, J. (2019). Big data management challenges in health research-a literature review. *Briefings in Bioinformatics*, *20*(1), 156-167. https://doi.org/10.1093/bib/bbx086

Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020). Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3655-3662. https://doi.org/10.1109/WACV45572.2020.9093435

Will McGugan. (2020, janvier). *Rich is a Python library for rich text and beautiful formatting in the terminal*. Récupérée 10 juin 2023, à partir de https://github.com/Textualize/rich

Willemink, M. J., Roth, H. R., & Sandfort, V. (2022). Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. *Radiology. Artificial Intelligence*, *4*(6), e210284. https://doi.org/10.1148/ryai.210284

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023, avril 27). Harnessing the Power of LLMs in Practice : A Survey on ChatGPT and Beyond. Récupérée 23 juin 2023, à partir de http://arxiv.org/abs/2304.13712

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. *npj Digital Medicine*, *5*(1), 1-9. https://doi.org/10.1038/s41746-022-00742-2

Zheng, G.-Y., Zeng, T., & Li, Y.-X. (2021). Application and prospect of cutting-edge information technology in biomedical big data. *Yi Chuan = Hereditas*, *43*(10), 924-929. https://doi.org/10.16288/j.yczz21-192

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020, janvier 8). Fine-Tuning Language Models from Human Preferences. https://doi.org/10.48550/arXiv.1909.08593

Université de Strasbourg

Corentin MEYER

École doctorale Sciences de la vie et de la santé | ED 414 Université de Strasbourg

Méthodes d'exploitation de données multimodales de patients par intelligence artificielle : application aux myopathies congénitales

Résumé

La croissance exponentielle des données générées par le séquençage, l'imagerie et les dossiers médicaux électroniques met en évidence le besoin d'outils pour l'exploitation des données biomédicales multimodales. L'objectif de ma thèse est de développer des méthodes basées sur l'intelligence artificielle pour explorer les données de patients atteints de myopathies congénitales, une famille de maladies génétiques rares et difficiles à diagnostiquer. Dans un premier temps, j'ai développé **IMPatienT** (*Integrated digital Multimodal PATIENt daTa*), une application web permettant d'annoter et d'explorer les rapports et les images histologiques des patients dont la base de données a été analysée par IA explicable (xAI). Ensuite, j'ai développé **NLMyo** (*Natural Language Myopathies*), un outil basé sur les modèles linguistiques de grande taille comme *GPT-3.5* et *LLaMA*. **NLMyo** permet d'anonymiser et d'extraire de l'information de comptes rendus médicaux, de faire de l'aide au diagnostic et de créer un moteur de recherche de patients. Enfin, j'ai développé **MyoQuant**, un outil utilisant des modèles d'IA pour quantifier automatiquement des marqueurs pathologiques dans les biopsies de fibres musculaires. **IMPatient**, **NLMyo** et **MyoQuant** sont disponibles de manière open-source et en version de démonstration en ligne.

Mots-clés : IA, quantification d'images, apprentissage profond, NLP, LLMs, myopathie congénitale, données biomédicales, rapports médicaux en texte libre, histologie, médecine translationnelle

Summary

The exponential growth of data generated by sequencing, imaging, and electronic medical records highlight the need for tools to exploit multimodal biomedical data. The objective of my thesis is to develop methods based on artificial intelligence to explore data from patients with congenital myopathies, a family of rare genetic diseases difficult to diagnose. First, I developed **IMPatienT** (Integrated digital Multimodal PATIENt daTa), a web application for annotating and exploring patient reports and histological images whose database has been analyzed by explainable AI (xAI). Then I developed **NLMyo** (Natural Language Myopathies), a tool based on large language models such as *GPT-3.5* and *LLaMA*. **NLMyo** allows anonymizing and extracting information from medical reports, to do diagnostic assistance and to create a patient search engine. Finally, I developed **MyoQuant** a tool using AI models to automatically quantify pathological markers in muscle fiber biopsies. **IMPatient**, **NLMyo** and **MyoQuant** are available as open-source and online demo versions.

Keywords: AI, image quantification, deep learning, NLP, LLMs, congenital myopathy, biomedical data, free text medical reports, histology, translational medicine