Intégration de connaissances médicales au sein d'un algorithme de classification automatique : application au codage du diabète

Arnaud Serret-Larmande[1], Jean-Baptiste Escudie[1], Catherine Duclos[12]

[1] HÔPITAL AVICENNE, Bobigny, France
arnaud.serret-larmande@aphp.fr
catherine.duclos@aphp.fr
jean-baptiste.escudie@aphp.fr

[2] INSERM UMRS 1142,
Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, Paris, France
catherine.duclos@aphp.fr

Résumé : La plus-value de l'adjonction de connaissances médicales sous formes de règles à un algorithme d'apprentissage automatique est évaluée ici au travers d'un cas d'étude : l'assignation de codes diagnostics de la Classification Internationale des Maladies (CIM-10) à des séjours de diabétologie à partir de documents textuels. La méthodologie hiérarchique développée ici entend simplifier la tâche de prédiction des algorithmes grâce à l'exploitation de la structure hiérarchique de la CIM-10. Les résultats montrent une augmentation des performances de prédiction de près de 10 points de F-score en moyenne pour la méthode hiérarchique comparativement à une méthode traitant chacun des codes de façon indépendante. L'utilisation à plus grande échelle d'une telle approche reste à explorer, et pourrait passer par l'exploitation de terminologies intégrant des relations conceptuelles plus détaillées que pour la CIM-10.

Mots-clés : Apprentissage automatique, expertise médicale, codage hospitalier, CIM-10

1 Introduction

À l'heure où les fruits des recherches en intelligence artificielle en médecine commencent à être utilisé dans la pratique courante (Abràmoff *et al.*, 2018) (Kalyanpur & Murdock, 2015), la question de la place du médecin et de son expertise médicale dans cette révolution en marche est amenée à se poser de façon continue dans un futur proche (Wartman & Combs, 2019). Malgré l'importance de la question, la difficulté pour intégrer cette expertise à la conception de systèmes d'apprentissage automatique rend le sujet délicat à traiter.

Les performances des applications basées sur de l'intelligence artificielle dans le domaine médical varient selon la tâche, et si en imagerie des algorithmes obtiennent d'ores-et-déjà des performances diagnostiques supérieures à celles de radiologues pour quelques cas d'usages (Liu *et al.*, 2018) (Steiner *et al.*, 2018), les applications à d'autres domaines médicaux n'ont pas encore montré d'amélioration substantielle, notamment dans le cas du codage hospitalier (Stanfill *et al.*, 2010 Nov Dec) (Sheikhalishahi *et al.*, 2019) (Topol, 2019). La complexité intrinsèque du domaine médical peut donc apparaître aujourd'hui comme rédhibitoire pour certaines tâches de classification. Ainsi, formaliser des connaissances médicales dans un format exploitable informatiquement pourrait représenter une réponse à ces difficultés.

Pour illustrer ce propos, nous avons souhaité évaluer l'apport de connaissances médicales formalisées à un algorithme de classification en comparant cette approche à un algorithme naïf, sur un cas d'usage précis : l'assignation de codes diagnostics de la Classification Internationale des Maladies (CIM-10) à des comptes rendus médicaux de diabétologie.

L'automatisation du codage des séjours hospitaliers par des algorithmes d'apprentissage est un champ de recherche important, tant du côté universitaire qu'industriel (Xie & Xing, 2018). Les enjeux sont en effet multiples, le codage des séjours hospitaliers n'étant plus seulement utilisé comme base du financement des établissements de soins, mais exploité de plus en plus fréquemment notamment pour l'évaluation de la qualité des soins, ou dans le cadre de la recherche biomédicale, favorisé notamment par les avancées récentes en intelligence

artificielle (Daien *et al.*, 2017) (Rajkomar *et al.*, 2018). Cependant, les performances des diverses approches réalisées pour assigner automatiquement des codes diagnostics aux séjours hospitaliers se sont révélées insuffisantes jusqu'à présent pour envisager leur utilisation en pratique courante.

Une approche naïve consiste à considérer les codes diagnostics issus de la CIM-10 comme un ensemble de labels indépendants, et a entraîner un algorithme évaluant la probabilité de chacun de ces codes d'être assigné à un séjour donné. Or la CIM-10 est organisée de façon hiérarchique en suivant une succession d'embranchements, depuis les chapitres généraux représentant des grands groupes nosologiques jusqu'aux codes diagnostics terminaux. Plusieurs codes peuvent ainsi partager une partie de leur signification respective en fonction de leur proximité dans cette arborescence. Quelques auteurs ont tenté d'exploiter cette structure hiérarchique de façon automatisée, par exemple via l'utilisation de réseaux de neurones convolutionnels dessinés pour apprendre la structure hiérarchique (Catling *et al.*, 2018), ou via une approche pas-à-pas descendante (Perotte *et al.*, 2014). Dans ces deux cas, l'utilisation de l'approche hiérarchique a permis d'améliorer les performances des modèles.

Nous avons souhaité ici évaluer une méthodologie mettant à profit la hiérarchie et les similitudes nosologiques existants parmis les codes diagnostics issus de la CIM-10 afin d'améliorer les performances de classification de sept algorithmes d'apprentissage automatique. Cette méthodologie développée spécifiquement pour cette étude sera dénommée méthode "hiérarchique", par opposition à l'approche naïve traitant l'ensemble des codes diagnostics comme indépendants, ci-après dénommée "indépendante".

2 Matériel et méthodes

2.1 Données et labels

Les données utilisées pour cette tâche de classification sont les comptes rendus médicaux, notes cliniques et ordonnance associés aux séjours du service de diabétologie de l'hôpital Avicenne (Bobigny, France), enregistrés entre le 1er septembre 2017 et le 1er février 2019. Le critère d'inclusion principal était la présence d'au moins un code CIM-10 appartenant aux sous-chapitres diabète de type 1 (E10.) ou 2 (E11.) en tant que diagnostic principal, diagnostic relié ou diagnostic associé secondaire. Cet ensemble de documents a été choisi pour deux principales raisons. Après application de ces critères de sélection, 1049 séjours ont été retenus.

Type de diabète		Type1	Type 2	
Insulino-dépendance		Oui	Oui	Non
Complications	Coma	E100	E1100	E1108
	Acidocétose	E101	E1110	E1118
	Rénale	E102	E1120	E1128
	Oculaire	E103	E1130	E1138
	Neurologique	E104	E1140	E1148
	Vasculaire	E105	E1150	E1158
	Autres précisées	E106	E1160	E1168
	Aucune complication	E109	E1190	E1198

Tableau 1 – Codes diagnostics des sous-chapitres diabète de type 1 et 2

Les labels retenus pour cette tâche de classification sont les codes issus de la CIM-10 adaptée par l'ATIH (Agence Technique de l'Information sur l'Hospitalisation) pour le PMSI français, appartenant à l'une des subdivisions des branches diabète de type 1 ou diabète de type 2 (Tableau 1). Après exclusion des codes interdits pour le codage des pathologies en lien avec le diabète d'après les recommandations de l'ATIH, 24 codes terminaux faisaient finalement partie des labels pouvant être assignés aux séjours hospitaliers.

Les séjours présentant des critères de mauvaise qualité de codage présence de codes mutuellement incompatibles (concommittance de 2 codes de diabète de type 1 et de type 2concernant le type du diabète, concommitance de 2 codes de diabète sans complication et avec complication, concommitance de 2 code de diabète insulino requérant et non insulino requérant), ou codes interdits pour le codage du diabète (omplications multiples et complications non précisées) ont été exclus. Après application de ces critères d'exclusions, 977 séjours ont finalement été retenus pour l'analyse (voir Figure 2).

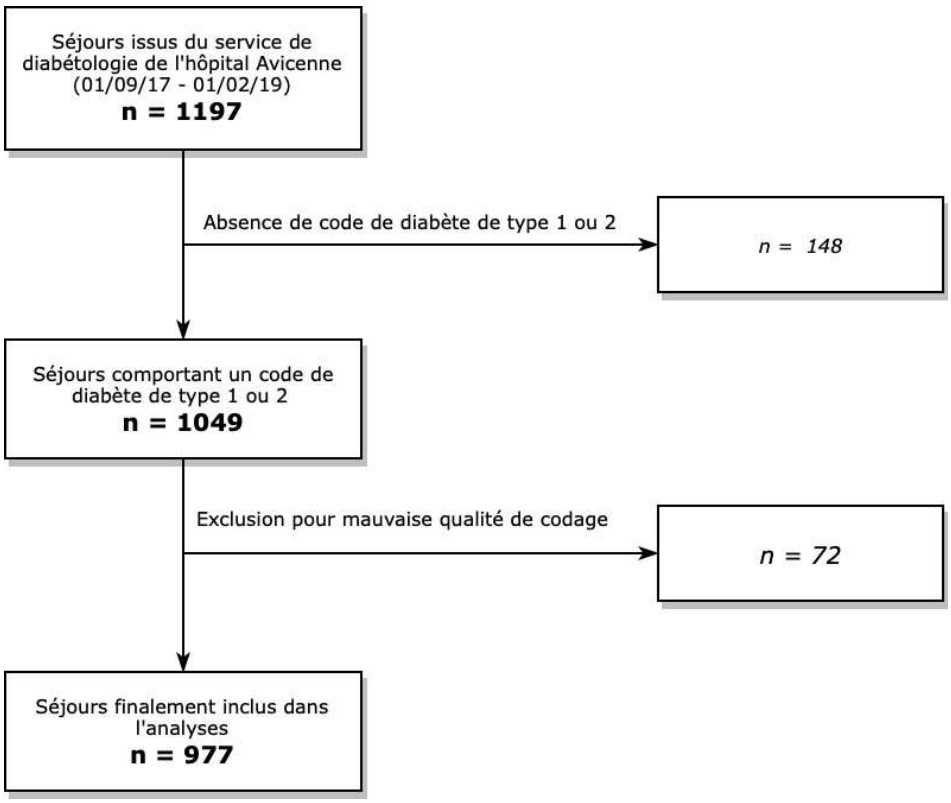

FIGURE *1 – Flow-chart : sélection des séjours*

2.2 Méthodologies de prédiction

L'assignation de codes CIM-10 à des comptes rendus de séjours hospitaliers correspond à une tâche de classification multi-catégorielle et multi-label : un séjour peut se voir attribuer un ou plusieurs codes diagnostics, et ceux-ci ne sont pas mutuellement exclusifs (par exemple l'ensemble de code [E101, E104, E105] peut être codé pour un même séjour) .

Concernant la méthodologie indépendante, le système devait être capable de prédire chacun des codes indépendamment les uns des autres. Nous avons ainsi entraîné un ensemble de classifieurs binaires, un pour chaque code présent dans le gold-standard de notre jeu de données. Les prédictions de ces classifieurs étaient *in fine* regroupées pour donner l'ensemble de codes assignés à un séjour.

Concernant la méthodologie hiérarchique, les labels terminaux ont été décomposés en plusieurs parties en fonction de leur signification, ce qui consistait ici à subdiviser chaque code en trois parties : une première pour le type du diabète, une deuxième pour les différentes complications liées au diabète et finalement une troisième pour la présence ou l'absence d'une requérance à l'insuline (voir Figure 2). Le système était conçu pour prédire spécifiquement chacune de ces sous-parties, à l'aide de multiples classifieurs binaires. Les prédictions étaient ensuite regroupées et les différents codes reconstitués à partir de la combinaison de chacune

de ces prédictions. Par exemple, si le système prédisait un diabète de type 2 (E11.), avec des complications oculaires (.2), rénales (.3) et vasculaires périphériques (.5) ainsi qu'une absence de requérance à l'insuline (..8), la combinaison de ces prédictions donnait l'ensemble de codes suivant : [E1128, E1138, E1158].

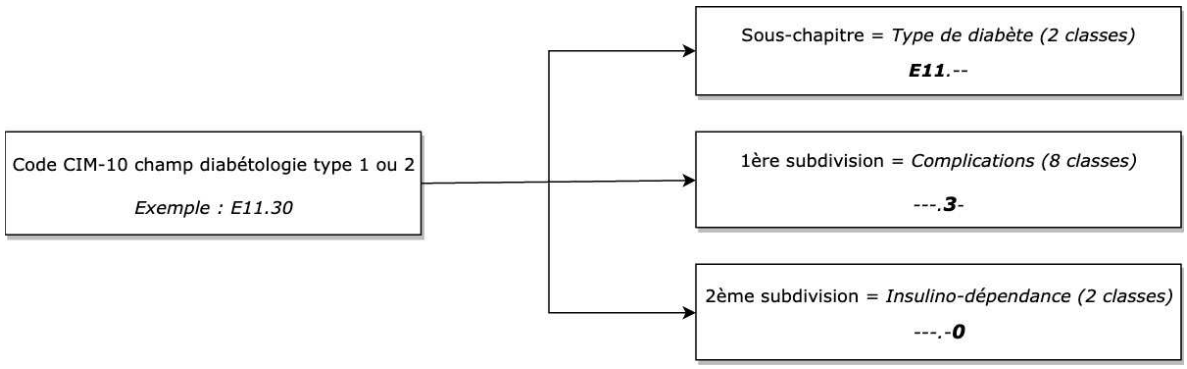

FIGURE *2 – Décomposition des codes terminaux par la méthodologie hiérarchique*

Cette décomposition de la tâche basée sur le sens médical des codes représentait l'expertise médicale apportée à l'algorithme : la décomposition des codes à prédire suivi d'une reconstruction à partir des prédictions permettait de réduire l'espace de prédictions possible et au passage de s'affranchir de la possibilité de prédiction de codes incompatibles. En effet, là où une méthodologie standard traitant chaque code de façon indépendante peut au final prédire des combinaisons de codes aberrantes pour un même séjour, telles que E10.1 (diabète de type 1 avec acidocétose) et E11.90 (diabète de type 2 insulino-dépendant sans complications), la reconstruction de codes à partir de prédictions intermédiaires permet d'éviter cet écueil. Secondairement, décomposer les codes de cette façon permettait d'obtenir plus d'exemples d'entraînement pour chacun des labels à prédire. En effet la décomposition en label intermédiaire permettait de regrouper tous les exemples positifs pour une même complication ou un même type de diabète ensemble, ce qui présentait un intérêt dans le cas de code décrivant des diagnostics dont la prévalence est très faible, et ce indépendemment de la taille du jeu d'entraînement. Par exemple, le code E1118 diabète de type 2 avec acidocétose non dépendant à l'insuline est peu fréquent de par la rareté même d'une telle pathologie, bien que les composants diabètes de type 2, acidocétose et dépendance à l'insuline ne soient pas spécifiquement rare pris isolément.

2.3 Algorithmes de classifications et pré-traitement des données

Les prédictions ont été réalisées à partir de l'ensemble des documents textuels disponibles pour un séjour. Chaque mot a été tokénizé et encodé selon son indice Term-Frequency/Inverse-Document-Frequency.

Le jeu de données a été divisé en 2 parties, avec 80% des données pour le jeu d'entraînement et 20% des données pour le jeu de test qui a servi à l'évaluation des performances.

Pour chacune des deux méthodologies présentées, les algorithmes d'apprentissage suivants ont été testés : Régression Logistique avec L1-pénalisation (Marquardt & Snee, 1975), Decision Tree, Random Forest, AdaBoost - Logistic Regression (Freund & Schapire, 1997), AdaBoost - Decision Tree (Freund & Schapire, 1997), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP). L'optimisation des hyperparamètres pour chacun de ces algorithmes a été réalisée selon l'approche *random search*, celle-ci pouvant converger vers un ensemble d'hyperparamètres optimal plus rapidement que la méthode *grid search* (Bergstra & Bengio, 2012). De plus un algorithme de classification naïf (Dummy Classifier) a été utilisé afin d'avoir une comparaison de base, celui-ci prédisant pour un séjour chaque label en fonction de la probabilité *a priori* de ce label (prévalence du label dans les données d'entraînement), indépendamment des documents textuels associés à ce séjour.

2.4 Evaluation

Les performances en matière de prédictions ont été évaluées selon trois métriques : la précision (ou valeur prédictive positive, rapport vrais positifs par nombre d'exemples prédits comme positifs), le rappel (ou sensibilité, rapport vrais positifs par nombre d'exemples réellement positifs) et le F-score (moyenne harmonique de la précision et du rappel). Ces métriques ont été retenues en se basant sur la littérature de référence (Koyejo *et al.*, 2014). Etant donné la nature multi-catégorielle et multi-label de la tâche de prédiction, un indicateur synthétique a été obtenu pour chacune des trois métriques par agrégation du nombre total de vrais positifs, faux positifs, vrais négatifs et faux négatifs pour chacun des labels – micro-averaging (Perotte *et al.*, 2014).

Les performances de prédiction ont été évaluées pour la prédiction des labels terminaux, à savoir les codes CIM-10 utilisés pour le codage des diagnostics, mais également pour la prédiction des labels intermédiaires construits pour l'implémentation de la méthode hiérarchique (et pouvant être déduits des prédictions de la méthodologie indépendante).

2.5 Logiciel

L'ensemble du code pour ce travail a été développé en python 3.6. Les librairies Scikit-learn 0.20 (Pedregosa *et al.*, 2011) et NLTK 3.4 (Bird *et al.*, 2009) ont été utilisées pour le pré-traitement des données, et Scikit-learn 0.20 pour l'implémentation des algorithmes et la recherche d'hyperparamètres.

3 Résultats

3.1 Description de la population

Au total, les 977 séjours inclus rassemblaient 3761 documents.

La répartition des labels est donnée par le tableau 2. La première ligne donne le nombre d'occurrences du code diagnostic dans l'ensemble du corpus, la deuxième la fréquence rapportée au nombre de séjours. Ainsi, la distribution des codes apparaît très déséquilibrée, avec une fréquence moyenne par code et par séjour de 8,5% (en excluant les codes complications "0", désignant les diabètes compliqués de coma, pris en charge en réanimation et non en service de diabétologie), mais pouvant aller de moins de 1% pour les codes E105 (diabète de type 1 avec complications vasculaires périphériques et E1118 (diabète sucré de type 2 non insulinotraité avec acidocétose), à plus de 35% pour le code E1120 (diabète de type 2 insulinotraité avec complications rénales).

Codes CIM-10	E100	E101	E102	E103	E104	E105	E106	E109	E1100	E1108	E1110	E1118
Nombre d'occurrences	0	39	40	68	52	1	21	60	0	0	27	1
Fréquence	0	0,04	0,04	0,07	0,05	0,	0,02	0,06	0	0	0,03	0
Codes CIM-10	E1120	E1128	E1130	E1138	E1140	E1148	E1150	E1158	E1160	E1168	E1190	E1198
Nombre d'occurences	343	31	315	25	326	37	24	6	179	22	110	30
Fréquence	0,35	0,03	0,32	0,03	0,33	0,04	0,02	0,01	0,18	0,02	0,11	0,03

Tableau 2 – Fréquence des différents codes par séjours (première ligne = nombre d'occurence dans corpus, deuxième ligne fréquence rapportée au nombre de séjours)

3.2 Prédictions des codes terminaux

Les performances de prédiction des labels terminaux sont résumées dans le tableau 3. La performance la plus élevée de prédiction globale est obtenue par la méthodologie hiérarchique avec pour classifieur la régression logistique pénalisée (F-score = 0,576). En utilisant la méthodologie indépendante, la performance la plus élevée est obtenue en utilisant un SVM pour estimateur, mais avec une performance relativement inférieure (F-score = 0,493). Pour

Algorithme	Méthodologie	F-score	Précision	Rappel
Logistic Regression	Hiérarchique	**0,576**	**0,587**	**0,565**
	Indépendante	0,481	0,451	0,515
Random Forest	Hiérarchique	0,574	**0,599**	0,550
	Indépendante	0,472	0,390	**0,597**
AdaBoost - Decision Tree	Hiérarchique	**0,566**	**0,574**	**0,559**
	Indépendante	0,467	0,445	0,491
SVM	Hiérarchique	**0,545**	0,551	**0,538**
	Indépendante	0,493	**0,560**	0,441
AdaBoost - Logistic Regression	Hiérarchique	**0,538**	**0,563**	0,515
	Indépendante	0,387	0,290	**0,582**
MLP	Hiérarchique	**0,511**	0,501	**0,521**
	Indépendante	0,411	**0,664**	0,297
Decision Tree	Hiérarchique	0,177	0,161	0,197
	Indépendante	**0,330**	**0,235**	**0,550**
Dummy Classifier	Hiérarchique	0,211	0,214	0,209
	Indépendante	0,212	0,202	0,224

Tableau 3 – Performance pour la prédiction des codes terminaux

chacun des algorithmes d'apprentissage testés, les performances semblent significativement meilleures avec l'utilisation de la méthodologie de prédiction dite hiérarchique (F-score augmenté de 9 points ou plus pour 5 des 7 algorithmes testés).

S'il apparaît une tendance nette à l'amélioration du F-score grâce à la méthodologie hiérarchique (hormis avec l'utilisation du Decision Tree), l'effet sur la précision et le rappel n'est pas uniforme et l'un ou l'autre de ces indicateurs peut être augmenté ou diminué selon les différents algorithmes.

Concernant la prédiction label par label les performances sont grandement influencées par la prévalence du code dans le jeu de donnée, mais on ne retrouve pas d'amélioration nette de la performance pour les labels peu fréquents en utilisant la méthodologie hiérarchique (résultats non montrés).

3.3 Performances dans la prédiction des codes intermédiaires

Les performances de prédictions pour les labels intermédiaires sont indiquées dans le tableau 4 (seul l'algorithme ayant les meilleurs performances de prédiction les labels terminaux est reporté). Les performances de prédiction apparaissent substantiellement améliorées pour chacun des trois labels intermédiaires, dans des proportions plus importantes pour le type de diabète et les différentes complications que pour l'insulino-requérance.

4 Discussion

Ainsi, la méthodologie hiérarchique implémentée ici a obtenu une augmentation moyenne de près de 10 points de F-score pour les algorithmes les plus performants.

Deux hypothèses théoriques semblaient susceptibles d'améliorer les performances de prédiction de la méthodologie hiérarchique : l'augmentation du nombre d'exemples d'entraînement par classifieur et la simplification de la tâche de prédiction. Les résultats code par code montraient que les codes les moins fréquents (donc les plus susceptibles de bénéficier d'une

Niveau	Algorithme	Méthodologie	F-score	Précision	Rappel
Complications	Logistic Regression	Hiérarchique	**0,696**	**0,709**	**0,682**
		Indépendante	0,618	0,606	0,629
	Dummy Classifier	Hiérarchique	0,360	0,364	0,356
		Indépendante	0,327	0,319	0,335
Insulino-dépendance	Logistic Regression	Hiérarchique	**0,803**	**0,795**	**0,810**
		Indépendante	0,717	0,665	0,778
	Dummy Classifier	Hiérarchique	0,627	0,607	0,647
		Indépendante	0,631	0,554	0,732
Type diabète	Logistic Regression	Hiérarchique	**0,903**	**0,903**	**0,903**
		Indépendante	0,779	0,794	0,765
	Dummy Classifier	Hiérarchique	0,704	0,704	0,704
		Indépendante	0,698	0,668	0,730

Tableau 4 – Performance pour la prédiction des codes intermédiaires

augmentation du nombre d'exemples d'entraînement) ne semblaient pas prédits avec une plus grande précision. Ainsi, l'augmentation globale des performances pourrait être attribuée au moins en partie à la simplification de la tâche de classification.

Un autre aspect bénéfique de cette implémentation hiérarchique pourrait être la plus grande proximité des codes prédits avec les codes réels. En effet, en regardant les performances dans la prédiction des labels intermédiaires, l'amélioration la plus importante est obtenue dans la prédiction du type de diabète (+12 points de F-score pour la régression logistique), c'est-à-dire pour le label intermédiaire représentant le noeud le plus en amont de l'arborescence de la CIM-10. Ainsi, dans le cas où la prédiction terminale s'avérerait fausse, les codes prédits pourraient se trouver tout de même plus proche du gold-standard, élément important dans l'optique d'une utilisation en pratique courante.

L'une des limitations de l'étude est l'utilisation de méthodes qui ne représentent pas l'état de l'art dans le champ du traitement automatique de la langue et de l'intelligence artificielle. En effet, l'encodage utilisé faisait appel à la méthode dite Bag-Of-Words (sac de mots), et non à l'utilisation d'embeddings qui constituent actuellement la méthode de référence pour l'encodage de texte en prenant en compte le contexte de chaque mot (Young *et al.*, 2018). Par ailleurs les algorithmes utilisés n'incluent pas de réseaux de neurones convolutionnels ou récurrents, qui sont les algorithmes de référence utilisés pour des tâches de prédictions à partir de documents textuels (Young *et al.*, 2017). La principale raison à cela était la taille limitée du corpus d'entraînement, celui-ci ayant été spécifiquement choisi en raison de la plus grande qualité des labels. Ce relativement faible nombre d'exemples ne permettait pas d'utiliser des modèles d'apprentissage profonds, dont les performances sur des corpus de tailles réduites sont notoirement limitées en raison du surapprentissage. Cependant, il nous semble que ces limitations ne remettent pas en question la validité des résultats présentés. D'une part, les performances d'algorithmes standards peuvent soutenir la comparaison dans certaines tâches, même comparés à des réseaux de neurones profonds (Rajkomar *et al.*, 2018), et d'autre part, en temps qu'étude de cas, nous souhaitions étudier l'impact de l'introduction de connaissances médicales pour améliorer les performances d'algorithmes d'apprentissage, indépendamment de leurs performances propres. Les résultats soutiennent ce constat : en effet, la variance en termes de qualité de prédiction dans l'éventail des différents algorithmes testés est moindre que la différence, pour chacun des algorithmes, entre les performances de la méthodologie hiérarchique et de la méthodologie indépendante (à l'exception de l'algorithme Decision Tree).

Une autre limitation de l'étude semble être la difficulté pour généraliser ce type de méthode à une tâche de prédiction plus large, idéalement incluant l'ensemble des codes diagnostics

possibles. S'il existe de nombreux chapitres et classes de la CIM-10 qui apparaissent éligibles à l'implémentation d'une méthodologie de prédiction hiérarchique, tels que la classe I60-I69 concernant les maladies cérébrovasculaires (combinaison d'un mécanisme physiopathologique et d'une localisation anatomique), ou encore la classe M15-M19 classant les arthroses (combinant localisations anatomiques et quantification de ces atteintes), il est vrai qu'une potentielle généralisation de cette approche nécessiterait un travail fastidieux pour implémenter une méthodologie hiérarchique pour un plus grand nombre de codes. D'autres approches ont tenté de mettre à profit la hiérarchie existante au sein de la CIM-10 en évitant d'avoir à implémenter un ensemble de règles manuelles (Catling *et al.*, 2018) (Perotte *et al.*, 2014), sans pour autant obtenir des performances suffisantes permettant d'envisager leur utilisation en pratique. Une solution à ce problème pourrait être l'utilisation de systèmes d'organisation des connaissances en santé plus spécialisés qui intègrent une composante relationnelle entre les pathologies de façon plus poussée que la CIM-10, permettant de concilier prise en compte des relations conceptuelles entres les pathologies et utilisation à grande échelle. On peut citer la terminologie SNOMED-CT qui implémente des relations entre phénomènes physiopathologiques et entités anatomiques, ou la 11ème version de la classification CIM (CIM-11) qui intègre la notion de parents multiples pour un même code diagnostic.

En conclusion, ce travail entend mettre en lumière les améliorations en terme de performances offertes lorsqu'un raisonnement médical est intégré dans la conception de systèmes de prédiction basés sur des algorithmes d'intelligence artificielle. L'utilisation d'une telle approche pour un ensemble de codes diagnostics exhaustif pourrait nécessiter de passer par des terminologies plus spécialisées intégrant des relations conceptuelles dans leur conception.

Références

ABRÀMOFF M. D., LAVIN P. T., BIRCH M., SHAH N. & FOLK J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, **1**(1), 39.

BERGSTRA J. & BENGIO Y. (2012). Random Search for Hyper-Parameter Optimization. p.̃25.

BIRD S., LOPER E. & KLEIN E. (2009). *Natural Language Processing With Python*. O'reilly media inc. edition.

CATLING F., SPITHOURAKIS G. P. & RIEDEL S. (2018). Towards automated clinical coding. *International Journal of Medical Informatics*, **120**, 50–61.

DAIEN V., KOROBELNIK J.-F., DELCOURT C., COUGNARD-GREGOIRE A., DELYFER M. N., BRON A. M., CARRIÈRE I., VILLAIN M., DAURES J. P., LACOMBE S., MARIET A. S., QUANTIN C. & CREUZOT-GARCHER C. (2017). French Medical-Administrative Database for Epidemiology and Safety in Ophthalmology (EPISAFE) : The EPISAFE Collaboration Program in Cataract Surgery. *Ophthalmic Research*, **58**(2), 67–73.

FREUND Y. & SCHAPIRE R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.

KALYANPUR A. & MURDOCK J. W. (2015). Unsupervised Entity-Relation Analysis in IBM Watson. (2015), 12.

KOYEJO O. O., NATARAJAN N., RAVIKUMAR P. K. & DHILLON I. S. (2014). Consistent Binary Classification with Generalized Performance Metrics. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. D. LAWRENCE & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 27*, p. 2744–2752. Curran Associates, Inc.

LIU Y., KOHLBERGER T., NOROUZI M., DAHL G. E., SMITH J. L., MOHTASHAMIAN A., OLSON N., PENG L. H., HIPP J. D. & STUMPE M. C. (2018). Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection. *Archives of Pathology & Laboratory Medicine*.

MARQUARDT D. W. & SNEE R. D. (1975). Ridge Regression in Practice. *The American Statistician*, **29**(1), 3–20.

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**, 2825–2830.

PEROTTE A., PIVOVAROV R., NATARAJAN K., WEISKOPF N., WOOD F. & ELHADAD N. (2014). Diagnosis code assignment : Models and evaluation metrics. *Journal of the American Medical*

Informatics Association : JAMIA, **21**(2), 231–237.

RAJKOMAR A., OREN E., CHEN K., DAI A. M., HAJAJ N., HARDT M., LIU P. J., LIU X., MARCUS J., SUN M., SUNDBERG P., YEE H., ZHANG K., ZHANG Y., FLORES G., DUGGAN G. E., IRVINE J., LE Q., LITSCH K., MOSSIN A., TANSUWAN J., WANG D., WEXLER J., WILSON J., LUDWIG D., VOLCHENBOUM S. L., CHOU K., PEARSON M., MADABUSHI S., SHAH N. H., BUTTE A. J., HOWELL M. D., CUI C., CORRADO G. S. & DEAN J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, **1**(1), 18.

SHEIKHALISHAHI S., MIOTTO R., DUDLEY J. T., LAVELLI A., RINALDI F. & OSMANI V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases : Systematic Review. *JMIR medical informatics*, **7**(2), e12239.

STANFILL M. H., WILLIAMS M., FENTON S. H., JENDERS R. A. & HERSH W. R. (2010 Nov-Dec). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA*, **17**(6), 646–651.

STEINER D., MACDONALD R., LIU Y., TRUSZKOWSKI P., HIPP J., GAMMAGE C., THNG F., PENG L. & STUMPE M. (2018). Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *The American Journal of Surgical Pathology*, **42**(12), 1636–1646.

TOPOL E. J. (2019). High-performance medicine : The convergence of human and artificial intelligence. *Nature Medicine*, **25**(1), 44.

WARTMAN S. A. & COMBS C. D. (2019). Reimagining Medical Education in the Age of AI. *AMA Journal of Ethics*, **21**(2), 146–152.

XIE P. & XING E. (2018). A Neural Architecture for Automated ICD Coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1066–1076, Melbourne, Australia : Association for Computational Linguistics.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. *arXiv :1708.02709 [cs]*.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, **13**(3), 55–75.