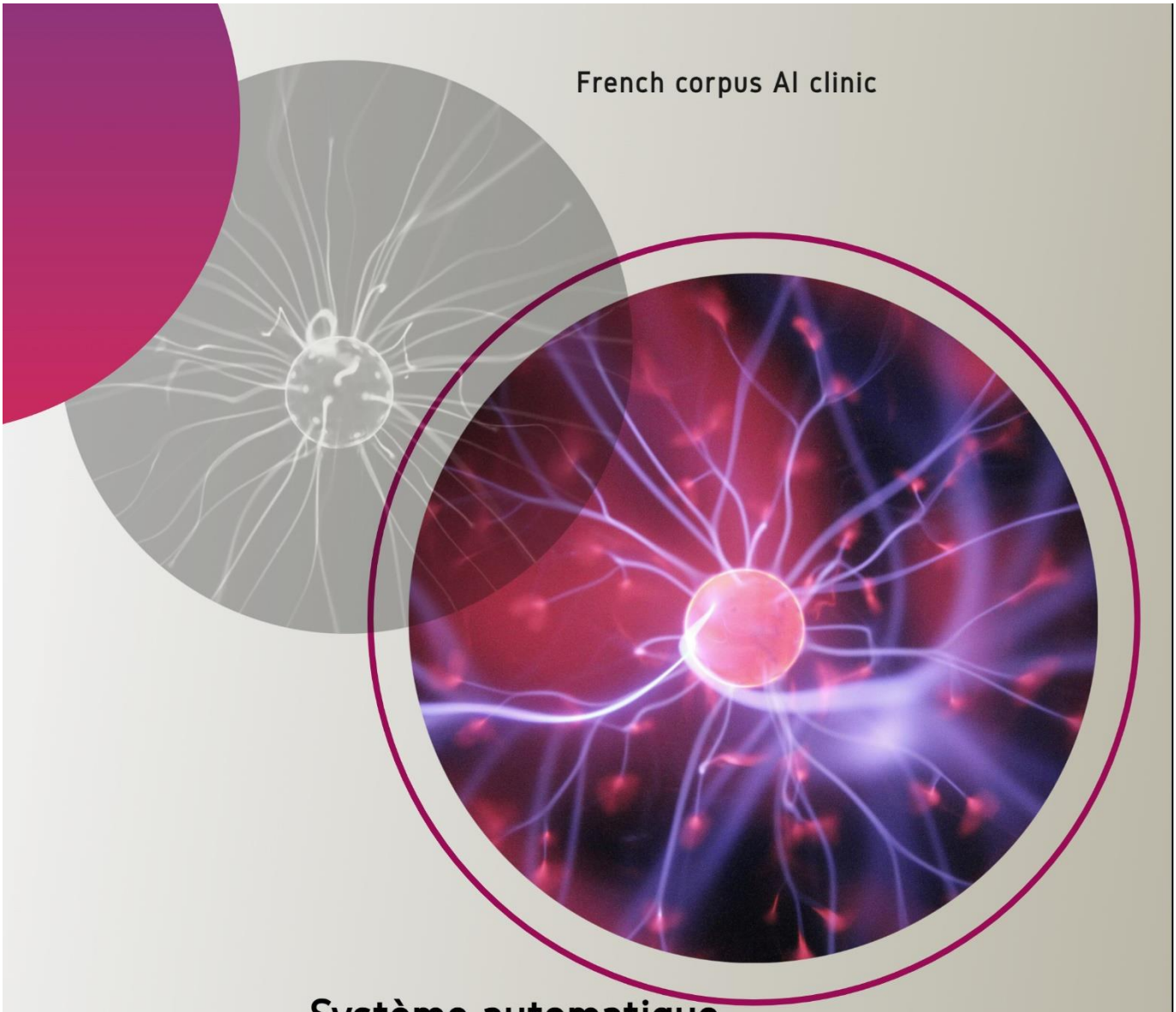
French corpus AI clinic



Système automatique de codage et de formation de la ICD-10

PGE 3

By

Ketsia Talotsing Magatsing

Table des matières

Abstrait

Background

Le code de la Classification internationale des maladies (CIM) est largement utilisé comme référence dans le système médical et à des fins de facturation. Cependant, la classification des maladies dans les codes de la CIM repose encore principalement sur la lecture par les humains d'une grande quantité de documents écrits comme base de codage. Le codage est à la fois laborieux et chronophage. Depuis la conversion de la CIM-9 en CIM-10, la tâche de codage est devenue beaucoup plus compliquée, et des approches liées à l'apprentissage profond et au traitement du langage naturel ont été étudiées pour aider les codeurs de maladies.

Objectif

Ce projet vise à construire un modèle d'apprentissage profond pour le codage de la CIM-10, où le modèle est destiné à déterminer automatiquement les codes de diagnostic et de procédure correspondants en se basant uniquement sur des notes médicales en texte libre afin d'améliorer la précision et de réduire l'effort humain.

Mots-clés : traitement automatique du langage naturel, apprentissage profond, classification internationale des maladies, réseau de neurones récurrents, classification de texte

Introduction

La Classification internationale des maladies (CIM) est une liste de classification médicale publiée par l'Organisation mondiale de la santé, qui définit l'univers des maladies, des troubles, des blessures et d'autres problèmes de santé connexes et la norme de classification du diagnostic. Depuis sa première publication en 1893, l'ICD est devenue l'un des index les plus importants dans les systèmes de gestion médicale, l'assurance maladie ou la recherche documentaire.

À l'heure actuelle, dans la plupart des établissements médicaux, les codes de la CIM-10 qui sont utilisés dans le cadre de la subvention de groupe liée au diagnostic pour les patients hospitalisés reposent principalement sur le codage manuel d'un groupe de codeurs de maladies agréés et professionnels au cas par cas, qui passent beaucoup de temps à lire une multitude de documents médicaux. D'autre part, d'autres cas, en particulier les patients ambulatoires, sont codés par les médecins.

Depuis la conversion de la CIM-9 en CIM-10 en 2014, Taïwan a utilisé la CIM-10 comme référence pour les subventions de groupe liées au diagnostic. Cependant, en raison de la complexité de la structure de la CIM-10 et des règles de codage, telles que l'ordre des codes, les critères d'inclusion et d'exclusion, et le nombre considérablement croissant de codes de la CIM-10, le travail de codage de la CIM-10 est devenu beaucoup plus laborieux et chronophage, même si un codeur de maladies ayant des compétences professionnelles prend environ 30 minutes par cas en moyenne. Selon l'analyse du Manuel de recherche sur l'informatique dans les soins de santé et la biomédecine, le coût de l'adoption du système CIM-10, y compris la formation des codeurs de maladies, des médecins et des utilisateurs de codes ; la perte de productivité initiale et à long terme chez les fournisseurs ; et la conversion séquentielle, est estimée à un coût unique de 1 millions de dollars US à 425,1 milliard de dollars US, en plus de 15 à 5 millions de dollars US par an en perte de productivité .

Des études antérieures avaient permis d'établir un modèle pour le système de la CIM-9. En 2008, Farkas et Szarvas ont utilisé une approche basée sur des règles en interrogeant d'autres outils de référence pour mettre en œuvre la tâche d'auto-codage de l'ICD. Cependant, par rapport à la CIM-9, la CIM-10 contient plus de 60 000 codes. La mise en place d'un système automatique basé sur des règles demande beaucoup de travail et de temps. De plus, l'ensemble des règles du système de la CIM-10 est compliqué, même pour les codeurs de maladies. Pour les raisons susmentionnées, des études récentes ont mis l'accent sur les approches liées à l'apprentissage profond et au traitement du langage naturel (NLP) ; par exemple, Zhang et al ont utilisé un réseau d'unités récurrentes fermées (GRU) avec une attention basée sur le contenu pour prédire les prescriptions de médicaments sur la base des codes de maladie, et Wang et al ont appliqué et comparé des techniques de NLP telles que Global Vectors (GloVe) dans une tâche de classification des données de dossiers de santé électroniques (DSE).

Ch. Blocks	Title	Ch. Blocks	Title
I. A00-B99	Certain infectious and parasitic diseases	XII. L00-L99	Diseases of the skin and subcutaneous tissue
II. C00-D48	Neoplasms	XIII. M00-M99	Diseases of the musculoskeletal system and connective tissue
III. D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	XIV. N00-N99	Diseases of the genitourinary system
IV. E00-E90	Endocrine, nutritional and metabolic diseases	XV. O00-O99	Pregnancy, childbirth and the puerperium
V. F00-F99	Mental and behavioral disorders	XVI. P00-P96	Certain conditions originating in the perinatal period
VI. G00-G99	Diseases of the nervous system	XVII. Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
VII. H00-H59	Diseases of the eye and adnexa	XVIII. R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
VIII. H60-H95	Diseases of the ear and mastoid process	XIX. S00-T98	Injury, poisoning and certain other consequences of external causes
IX. I00-I99	Diseases of the circulatory system	XX. V01-Y98	External causes of morbidity and mortality
X. J00-J99	Diseases of the respiratory system	XXI. Z00-Z99	Factors influencing health status and contact with health services
XI. K00-K93	Diseases of the digestive system	XXII. U00-U99	Codes for special purposes

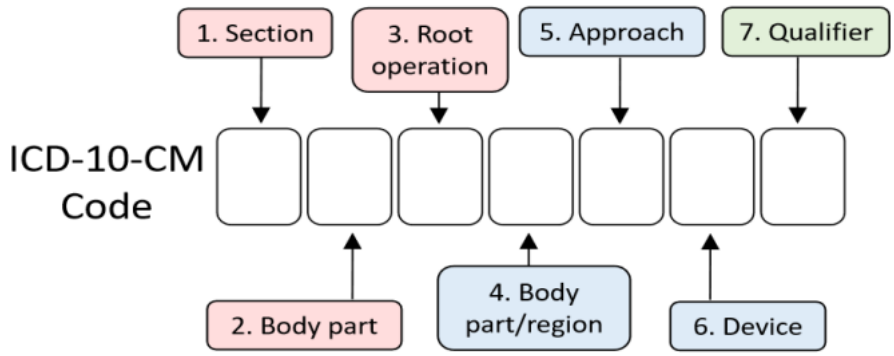


Figure 1: The ICD-10-CM structure.

Matériels et méthodes

Cette section décrit la collecte des données et details l'approche proposé

Description des données et Préprocessing

Les données présentées par l'entreprise UIZ.CARE était d'environ 1000 lignes pour plus de 60000 codes à prédire. De même les informations présentes sur la base de données ne correspondaient pas à notre attente notamment au niveau du diagnostique ou on avait plutôt la description universelle du code ICD au lieu d'un diagnostic fait en bonne et due forme.

Au vue de tout cela, nous nous sommes penchés à la recherche de nouvelles données qui peuvent mieux correspondre à nos attentes, nous sommes tombés sur une base de données portant sur les codes ICD et de plus en Français sur Kaggle. On avait alors une base de training (181000 lignes), de test et validation.

Nos données se présentaient comme suite :

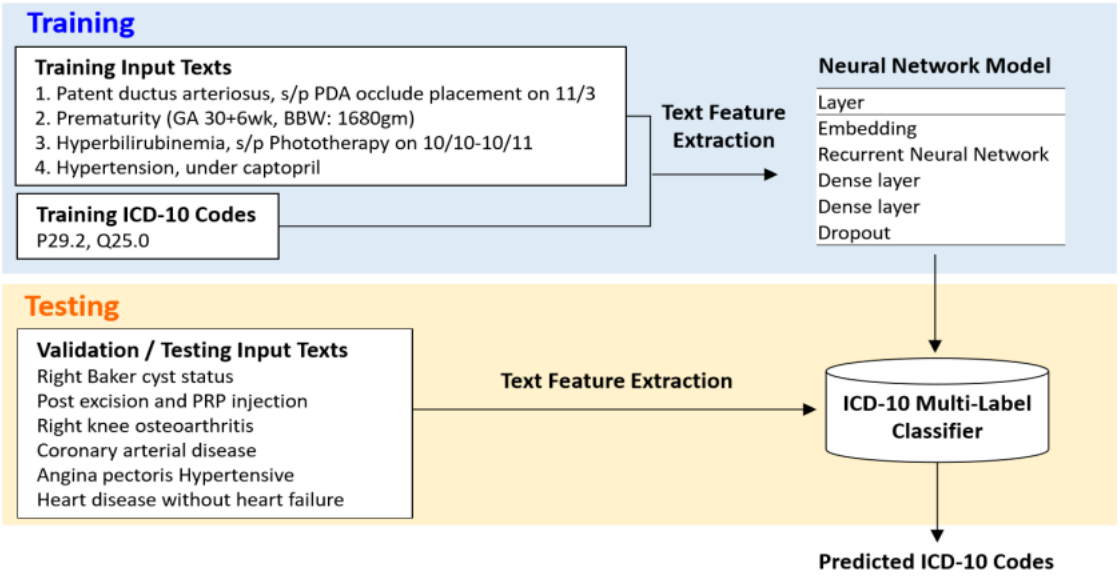
- RawText : qui représente le diagnostic fait sur le patient en texte libre e
- ICD10 : qui représente au code ICD correspondant

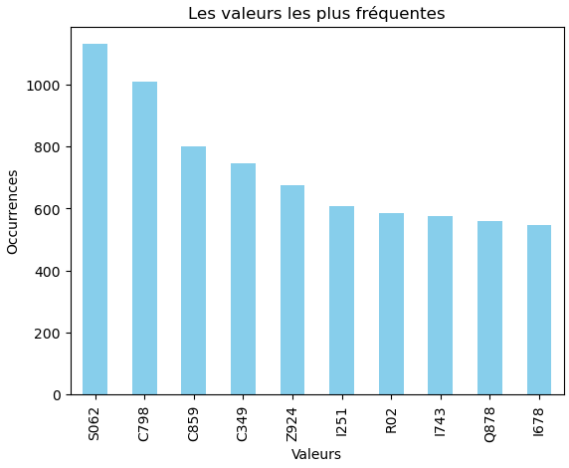
	RawText	ICD10
0	Thrombose veineuse profonde cuisse gauche	I802
1	Hémiplégie post-traumatique	S099
2	Masculinisation avec hyperplasie surrénale	E250
3	Hyperammoniémie cérébrale	E722
4	Fistule artérioveineuse congénitale périphériq...	Q257

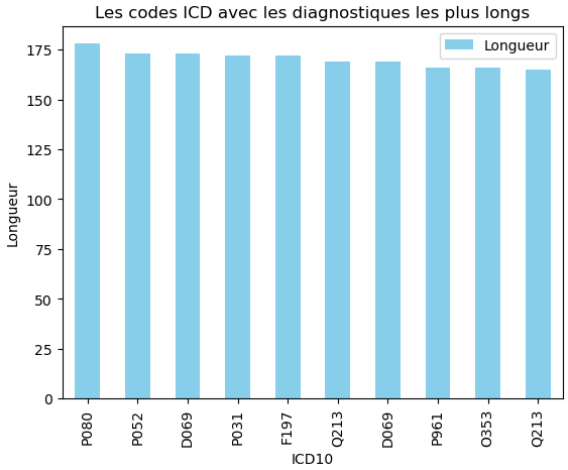
Extraction de caractéristiques à partir des diagnostics

Afin de simuler le travail des codeurs dans les hôpitaux, notre objectif est de construire un modèle qui prédit les codes CIM-10 en fonction des données textes de forme libre. Dans notre modèle, nous appliquons d'abord un prétraitement de base méthodes via NLTK, puis construisez un modèle de réseau neuronal pour apprendre les fonctionnalités des textes saisis. Le prétraitement la procédure comprend la vérification orthographique, la conversion en minuscules, suppression des mots vides, tokenisation et suppression des mots rares mots. Les données prétraitées sont ensuite divisées en formations et validation définie par la bibliothèque Scikit-Learn.

Dans le modèle de réseau neuronal, la première couche est le mot intégration layer, qui est le nom collectif d'un ensemble de modélisation de langage et présentent des techniques d'apprentissage en NLP où des mots ou des phrases du vocabulaire sont mappés en vecteurs de nombres réels. Nous codons ensuite chaque mot tokenisé dans son intégration de mots basé sur word2vec et GloVe, compte tenu de leur capacité de capturer la sémantique et la syntaxe dans les vecteurs.







Dans notre travail, nous obtenons les meilleures performances en utilisant l'intégration à 300 dimensions pour représenter les mots dans l'ensemble des documents. Dans l'apprentissage du modèle word2vec, nous utilisons un paramètre de taille de fenêtre de 10, ce qui indique que la distance maximale entre le mot actuel et le mot prédit se situe à l'intérieur d'une phrase. Le paramètre de comptage minimal est de 5, ce qui indique que le modèle ignore tous les mots dont la fréquence totale est inférieure à 5. Le paramètre d'échantillonnage est de 0,1, ce qui signifie que le seuil de configuration des mots à haute fréquence est échantillonné de manière aléatoire. Après l'entraînement, nous disposons de la couche d'intégration des mots, qui est la couche supérieure de notre modèle de réseau neuronal pour transformer toutes nos données de texte libre en format vectoriel, et notre modèle apprend alors les informations cachées dans les documents.

Dans notre travail, nous obtenons les meilleures performances en utilisant des encastrements à 300 dimensions pour représenter les mots dans l'ensemble des documents. Dans l'apprentissage du modèle word2vec, nous utilisons un paramètre de taille de fenêtre de 1, ce qui indique que la distance maximale entre le mot actuel et le mot prédit se situe dans une phrase. Le paramètre de comptage minimal est de 5, ce qui indique que le modèle ignore tous les mots dont la fréquence totale est inférieure à 5. Le paramètre d'échantillonnage est de 0,1, ce qui signifie que le seuil de configuration des mots à haute fréquence est aléatoirement

échantillonné vers le bas. Après l'entraînement, nous disposons de la couche d'intégration des mots, qui est la couche supérieure de notre modèle de réseau neuronal, pour transformer toutes nos données de texte libre en format vectoriel, et notre modèle apprend alors les informations cachées dans les documents.

Deep Neural Nerwork Model

L'apprentissage supervisé permet d'apprendre une fonction qui associe une entrée à une sortie sur la base des paires entrée-sortie. Nous devons préparer l'ensemble de données avec des données de formation étiquetées. Dans cette recherche, nos données ont été étiquetées au préalable. Chacune des données en texte libre comportait une paire de codes CIM-10 en tant qu'étiquette. Notre modèle de réseau neuronal a analysé les données de texte libre en entrée pour apprendre une fonction de mise en correspondance qui pourrait mettre en correspondance les données de texte libre avec les multiples codes CIM-10 corrects. À l'instar du concept d'apprentissage humain, notre modèle peut observer les données d'entrée et corriger l'idée que l'on se fait des données d'entrée à l'aide de l'étiquette, pour finalement comprendre la relation entre les données d'entrée et l'étiquette de sortie. La structure de notre modèle est un réseau neuronal à quatre couches, comme le montre la figure 3. La première couche est la couche d'intégration des mots, qui transforme le texte libre en vecteurs de mots. La deuxième couche est une couche d'unité récurrente bidirectionnelle (GRU). La GRU est un réseau neuronal récurrent doté de mécanismes de gating, ce qui permet de résoudre le problème de gradient de disparition qui se pose parfois avec les réseaux neuronaux récurrents standard. La GRU nécessite également moins de temps de calcul que la mémoire à long terme (LSTM). Les couches restantes sont deux couches denses avec une unité linéaire rectifiée (ReLU) et une sigmoïde comme fonction d'activation séparée, les couches denses finales devant produire le vecteur avec la dimension que nous prévoyons de prédire. Dans le cas de la classification en 21 catégories, la CIM-10 compte 21 catégories au total, de sorte que la couche dense finale doit produire un vecteur à 21 dimensions, où chaque dimension indique la probabilité d'association d'un code. Dans le cas de la classification de l'ensemble des étiquettes, la taille de la sortie doit être égale à la quantité d'étiquettes, c'est-à-dire la couche dense 2 dans le tableau 4. Le tableau 4 présente les paramètres de l'abandon et les quatre couches de notre modèle.

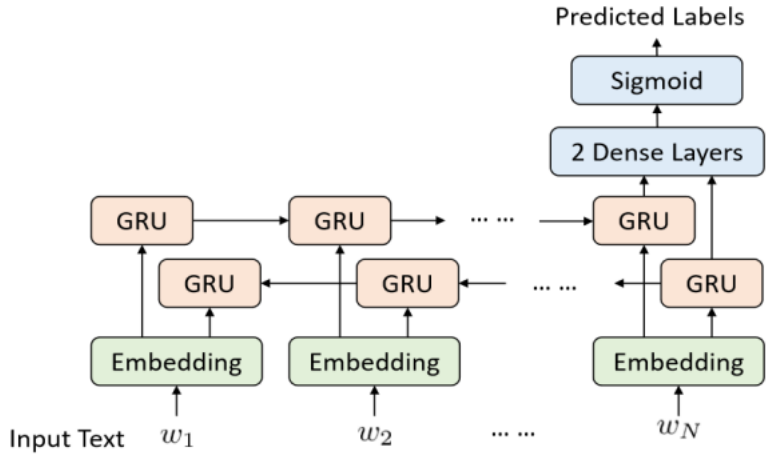


Figure 3: Neural network model structure used in this paper.

Hyper paramètres	Tailles
Embedding layer	300
Bidirectional GRU layer	256
Dense layer 1	700
Dense layer 2	14602
Dropout	0.2

Résultats et Discussion

Pour évaluer les performances de notre modèle, nous utilisons le score F1 comme mesure d'évaluation. Le score F1 est la moyenne harmonique du rappel et de la précision. La précision et le rappel peuvent évaluer les performances du modèle avec des faux positifs et des faux négatifs. Nous pensons donc que le score F1, qui équilibre les deux mesures, est adapté à notre objectif. Nous examinons ci-dessous trois paramètres expérimentaux pour la prédiction de diagnostic.

Tout d'abord, nous classons les codes en fonction des chapitres de la CIM. Dans la CIM-10 CM, il y a 22 blocs pour subdiviser les codes dans un format de 3 caractères. Chaque bloc a un titre qui présente la maladie. Par exemple, A00-B99 concerne « certaines maladies infectieuses et parasitaires ». Étant donné que les blocs U00-U99, qui concernent les « Codes à usage spécial », ne sont pas liés à des maladies, notre modèle ne prédit pas les codes CIM-10 entre U00 et U99. Par conséquent, notre modèle ne prédit que 21 catégories dans la CIM-10.

Résultat du modele

Les résultats du modèle ont démontré une amélioration significative au fil de l'entraînement sur quatre epochs successives. La perte initiale, élevée à 6.13, a connu une réduction drastique pour atteindre 1.22, indiquant une convergence progressive du modèle vers des prédictions plus précises. Parallèlement, l'accuracy a suivi une trajectoire ascendante, passant de 0.17 à 0.70, témoignant ainsi d'une meilleure justesse dans les prédictions effectuées par le modèle.

En ce qui concerne l'évaluation sur l'ensemble de validation, des résultats similaires ont été observés avec une diminution de la perte de 4.36 à 2.82 et une augmentation de la précision de 0.34 à 0.54. Ces améliorations soutiennent l'idée que le modèle a réussi à généraliser ses connaissances sur des données inconnues, bien qu'il reste crucial de tester sa capacité à généraliser sur un ensemble de données encore plus varié pour confirmer sa robustesse.

Il est important de noter que l'entraînement du modèle a nécessité une durée significative, d'environ 1h45 à 1h50 par epoch. Cette durée pourrait indiquer une complexité élevée du modèle ou la manipulation d'un volume conséquent de données.

En somme, ces résultats encourageants suggèrent un apprentissage efficace du modèle. Toutefois, une évaluation approfondie et des tests supplémentaires sur des données diverses sont essentiels pour valider pleinement sa capacité à généraliser et à produire des prédictions précises dans des contextes réels.

```
Epoch 1/10
4545/4545 [==============================] - 6457s 1s/step - loss: 6.1268 - accuracy: 0.1705 - val_loss: 4.3625 - val_accuracy: 0.3350
Epoch 2/10
4545/4545 [==============================] - 6421s 1s/step - loss: 3.1193 - accuracy: 0.4534 - val_loss: 3.2350 - val_accuracy: 0.4748
Epoch 3/10
4545/4545 [==============================] - 6350s 1s/step - loss: 1.8789 - accuracy: 0.6032 - val_loss: 2.8176 - val_accuracy: 0.5389
Epoch 4/10
3818/4545 [========================>.....] - ETA: 16:21 - loss: 1.2235 - accuracy: 0.7027
```

Conclusion

Ce projet de modélisation et de prédiction d'étiquettes ICD10 à partir de données textuelles a été une exploration enrichissante dans le domaine de l'intelligence artificielle appliquée à la santé. À travers ce parcours, plusieurs éléments clés ont été mis en lumière.

Tout d'abord, la préparation des données s'est avérée être une phase cruciale, impliquant le nettoyage, la tokenisation et la vectorisation des textes bruts. La détermination du vocabulaire, de la taille des séquences et la transformation des étiquettes ICD10 en encodage adéquat ont été des étapes fondamentales pour la formation d'un modèle pertinent.

L'architecture du modèle, basée sur des couches d'embedding, de GRU bidirectionnelles et de couches denses, a présenté une évolution progressive et des améliorations notables au fil de l'entraînement. Les résultats obtenus ont témoigné d'une augmentation significative de la précision et d'une réduction conséquente de la perte, reflétant l'apprentissage et la capacité du modèle à extraire des schémas significatifs à partir des données.

Néanmoins, des défis ont émergé, notamment la durée d'entraînement élevée, suggérant une complexité probable du modèle ou une manipulation d'un volume important de données. De plus, des analyses plus approfondies sur la généralisation du modèle et des tests sur des ensembles de données plus diversifiés sont nécessaires pour évaluer pleinement sa performance dans des cas réels et variés.

En résumé, ce projet a permis d'explorer et de mettre en œuvre différentes techniques d'apprentissage automatique pour la classification de données médicales. Malgré les défis rencontrés, les résultats encourageants soulignent le potentiel des modèles d'apprentissage profond dans le domaine de la santé, tout en soulignant la nécessité continue de recherches approfondies pour garantir des performances fiables et généralisables.