

---

## **Rapport de Projet (Phase 2) : Développement d'une IA pour l'Attribution Automatique des Codes ICD dans les Dossiers Médicaux**

---

### **Introduction**

Au cours du second semestre de notre initiative visant à transformer le processus d'attribution des codes ICD, notre équipe s'est concentrée sur le peaufinage et l'optimisation de notre solution d'IA. Face aux défis méthodologiques et techniques rencontrés initialement, nous avons exploré des stratégies innovantes pour améliorer la précision de notre modèle et faciliter son intégration dans les flux de travail médicaux. Cette phase du projet a été marquée par une collaboration étroite avec les professionnels de santé, une exploration approfondie des données médicales et le développement d'une interface utilisateur intuitive.

### **Création de Données et Data Augmentation**

#### **Collecte et Prétraitement des Données**

La collecte de données a été une entreprise ambitieuse, visant à compiler un ensemble de données représentatif et varié. Notre approche a impliqué la collaboration avec des institutions médicales pour accéder à une vaste collection de dossiers médicaux. Le prétraitement a joué un rôle crucial dans la conversion de ces données brutes en un format exploitable pour l'entraînement du modèle. Ce processus a inclus la normalisation du texte, la suppression des informations redondantes ou non pertinentes, et la sécurisation des données pour respecter la confidentialité des patients.

#### **Stratégies de Data Augmentation**

Reconnaissant les limitations de notre jeu de données initial, nous avons adopté des techniques de data augmentation pour enrichir notre ensemble de données. Ces méthodes ont non seulement augmenté la quantité de données disponibles mais ont également introduit une diversité linguistique et contextuelle, essentielle pour entraîner un modèle robuste et adaptable.

### **Défis et Solutions dans le Développement des Modèles**

#### **Surmontée de l'Overfitting**

L'overfitting a émergé comme un défi significatif, compromettant la capacité de notre modèle à généraliser à partir de données inédites. Pour y remédier, nous avons expérimenté avec différentes architectures de réseau neuronal, augmenté la régularisation, et ajusté le taux d'abandon. Ces ajustements ont contribué à améliorer la généralisation du modèle sans sacrifier sa performance sur l'ensemble d'entraînement.

#### **Optimisation des Hyperparamètres**

L'optimisation des hyperparamètres a été une étape cruciale pour atteindre une performance optimale. En utilisant des techniques comme la recherche par grille et l'optimisation

bayésienne, nous avons systématiquement exploré l'espace des hyperparamètres pour trouver la configuration idéale qui maximise la précision de notre modèle.

### Développement et Test de l'Interface Utilisateur

L'interface utilisateur, développée avec TKinter, a été conçue pour être à la fois simple et fonctionnelle, permettant aux utilisateurs de saisir facilement des diagnostics médicaux et de recevoir les codes ICD correspondants. Des tests d'utilisabilité ont été menés pour s'assurer que l'interface répondait aux besoins des utilisateurs finaux, entraînant plusieurs itérations de conception pour améliorer l'expérience utilisateur.

### Résultats, Évaluation et Perspectives

#### Performance du Modèle

Notre modèle a atteint une précision impressionnante de 86%, marquant une avancée significative par rapport aux benchmarks initiaux. Cette amélioration est attribuable à notre approche rigoureuse en matière de prétraitement des données, de data augmentation, et d'optimisation des modèles.

#### Évaluation et Retour d'Expérience

L'évaluation de notre système par des professionnels de la santé a révélé une forte appréciation de sa précision et de sa facilité d'utilisation. Le feedback recueilli a souligné l'importance de notre travail dans la réduction de la charge de travail des codeurs médicaux et dans l'amélioration de l'efficacité des processus de facturation.

#### Perspectives Futures

Fort de ces résultats prometteurs, nous envisageons d'élargir notre collaboration avec d'autres institutions médicales pour enrichir davantage notre jeu de données. Parallèlement, nous explorerons des technologies émergentes et des architectures de modèles plus avancées pour améliorer encore la précision et la rapidité de notre système. L'expansion de notre base de données pour inclure des langues et dialectes additionnels est également envisagée, dans le but de rendre notre solution accessible à une audience mondiale.

#### Impact Sociétal et Éthique

Notre projet s'inscrit dans une démarche résolument tournée vers l'amélioration de la qualité des soins de santé et l'efficience des systèmes médicaux. En automatisant le processus de codage ICD, nous contribuons à minimiser les erreurs humaines, à accélérer les processus administratifs et, in fine, à améliorer l'accès aux soins pour les patients. Toutefois, nous sommes pleinement conscients des implications éthiques liées à l'utilisation de l'IA dans le domaine de la santé, notamment en matière de confidentialité des données et de prise de décision automatisée. À cet égard, nous veillons à ce que notre modèle soit transparent, équitable et conforme aux réglementations en vigueur.

#### Collaboration Interdisciplinaire

La réussite de ce projet est le fruit d'une collaboration interdisciplinaire impliquant des informaticiens, des médecins, des codeurs médicaux et des spécialistes en éthique. Cette synergie entre différentes expertises a été cruciale pour adresser de manière holistique les défis techniques, médicaux et éthiques rencontrés. À l'avenir, nous prévoyons de renforcer ces

collaborations, notamment en intégrant des psychologues pour étudier l'acceptabilité de l'IA par les professionnels de santé et les patients.

## Développements Futurs

### Intégration dans les Systèmes d'Information Hospitaliers

L'une de nos principales ambitions est d'intégrer notre système d'IA directement dans les systèmes d'information hospitaliers (SIH) existants. Cela nécessitera un travail de développement supplémentaire pour assurer la compatibilité et la sécurité des données, mais représente une étape clé pour faciliter l'adoption de notre solution dans le quotidien des professionnels de santé.

### Extension des Capacités du Modèle

Nous envisageons également d'étendre les capacités de notre modèle pour couvrir non seulement les codes ICD mais aussi les recommandations de traitements et de diagnostics complémentaires. Cela pourrait transformer notre système en un outil d'aide à la décision médicale complet, augmentant son utilité pour les professionnels de santé.

## Conclusion

La phase 2 de notre projet a été une période d'apprentissage intense et de progrès significatifs. Malgré les défis rencontrés, les résultats obtenus témoignent de l'efficacité de notre approche et ouvrent la voie à des applications prometteuses de l'IA dans le domaine médical. Alors que nous nous tournons vers l'avenir, notre équipe reste dédiée à l'amélioration continue de notre système, avec l'ambition de contribuer à une révolution technologique bénéfique pour les systèmes de santé du monde entier.

## Remerciements

Nous tenons à exprimer notre profonde gratitude à tous ceux qui ont contribué à ce projet, depuis les institutions partenaires jusqu'aux individus qui ont partagé leurs connaissances et leur temps. Leur soutien a été indispensable à nos succès et continue d'inspirer notre engagement envers l'excellence et l'innovation.

## Notice Explicative du Dossier de Projet

Le dossier de projet contient tous les éléments nécessaires à la compréhension et à l'évaluation de la deuxième phase du développement de l'IA pour l'attribution automatique des codes ICD. Voici le détail des fichiers et dossiers inclus :

- **data** : Un dossier compressé contenant l'ensemble des jeux de données utilisés pour l'entraînement, la validation et le test du modèle d'IA. Ce dossier est essentiel pour examiner la qualité, la structure et la diversité des données sur lesquelles le modèle a été formé.
- **data\_augmentation\_with\_embedding** : Un fichier Jupyter Notebook qui détaille les techniques de data augmentation utilisées, y compris l'embedding de mots. Ce document explique comment les données ont été enrichies pour améliorer la performance du modèle.

- **ICD\_CODE\_Rapport\_phase1** : Un fichier PDF qui contient le rapport de la première phase du projet. Il sert de base de référence pour évaluer les progrès réalisés au cours de la deuxième phase.
- **LeModel.h5** : Le fichier de sauvegarde du modèle entraîné en format H5. Il inclut l'architecture du modèle ainsi que les poids des neurones après l'entraînement, permettant ainsi une réutilisation directe du modèle ou une continuation de l'entraînement.
- **NN\_Code\_ICD10** : Un fichier Jupyter Notebook qui présente le code source pour la définition, l'entraînement et l'évaluation du modèle de réseau de neurones utilisé pour la classification des codes ICD10.
- **Présentation\_ICD\_CODE** : Un fichier PowerPoint de présentation qui résume le projet, ses objectifs, méthodes, résultats et conclusions. Ce fichier est prévu pour les communications avec les parties prenantes et la diffusion des résultats du projet.
- **visualisation\_modelICD** : Un fichier Jupyter Notebook contenant des visualisations des résultats d'entraînement du modèle, telles que des graphiques de la loss et de l'accuracy. Ces visualisations aident à comprendre la performance et la convergence du modèle au fil des epochs.
- **execution\_model** : Un fichier Jupyter Notebook qui illustre comment exécuter le modèle pour faire des prédictions à partir de nouvelles entrées. Ce fichier sert de guide pratique pour utiliser le modèle dans un contexte opérationnel.

Chaque composant de ce dossier est conçu pour être à la fois autonome et partie intégrante d'un ensemble cohérent, facilitant ainsi la compréhension globale du projet et permettant d'apprécier les efforts déployés ainsi que les avancées technologiques réalisées durant cette phase de développement.