Prédiction de la trajectoire du patient : Intégration des notes cliniques aux transformers

Sifal Klioui*, Sana Sellami*, Youssef Trardi*

*Aix-Marseille Univ, LIS, CNRS Marseille, France
sifal.klioui@etu-univ-amu.fr
sana.sellami@univ-amu.fr
youssef.trardi@univ-amu.fr

Résumé. La prédiction des trajectoires de maladies à partir des dossiers médicaux électroniques (DME) pose des défis liés à la non-stationnarité des données, à la granularité des codes médicaux et à l'intégration de données multimodales. Les DME combinent données structurées, comme les codes de diagnostic, et données non structurées, telles que les notes cliniques, souvent sous-exploitées malgré leur richesse informative. Pour remédier à ces limites, nous proposons d'intégrer les notes cliniques non structurées dans des modèles de prédiction séquentielle des maladies basés sur les modèles Transformers, améliorant ainsi la précision des prédictions de diagnostics futurs. Les expérimentations menées sur les ensembles de données MIMIC-IV ont montré que l'approche proposée offre de meilleures performances en comparaison des modèles traditionnels reposant uniquement sur des données structurées.

1 Introduction

En santé, la croissance exponentielle des Dossiers Médicaux Électroniques (DME) a révolutionné les soins aux patients tout en posant de nouveaux défis. Les professionnels de santé interagissent désormais fréquemment avec des dossiers médicaux s'étendant sur plusieurs décennies, devant traiter et analyser cette vaste quantité d'informations pour prendre des décisions éclairées sur l'état de santé futur des patients. Cette évolution a accéléré le développement de systèmes automatisés pour prédire les diagnostics futurs à partir des données médicales passées, devenant ainsi un élément clé de la médecine personnalisée et proactive.

Les techniques d'apprentissage automatique, en particulier l'apprentissage profond, ont connu un essor croissant en médecine (Egger et al., 2022), grâce à leur capacité à s'adapter et à leurs bons résultats. En imagerie médicale, par exemple, les modèles d'apprentissage profond ont atteint un niveau de performance élevé dans la prédiction des diagnostics médicaux, parfois comparable, voire supérieur, à celui des experts humains. Ces résultats ont conduit les chercheurs à appliquer des techniques similaires à la tâche de prédiction séquentielle des maladies (Choi et al., 2016a; Rodrigues-Jr et al., 2021; Shankar et al., 2023), où l'objectif est de prédire le diagnostic d'un patient lors de sa prochaine visite (N+1) en s'appuyant sur le contenu de ses visites précédentes (N). Cependant, la modélisation des trajectoires des patients à partir des

données des DME présente des défis uniques : (i) la non-stationnarité des données des DME qui entraîne des variations dans les données, ce qui limite la généralisabilité des modèles, (ii) la granularité élevée des codes médicaux (par exemple, plus de 70 000 dans la Classification Internationale des Maladies, 10e Révision, Modification Clinique (ICD-10-CM [1])) qui rend difficile l'exploration et l'utilisation de ces codes par les modèles de prédiction, (iii) Les dépendances à long terme dues au traitement de longues séquences de données représentent une tâche difficile pour les modèles traditionnels de réseaux de neurones récurrents (RNN), (iv) l'intégration de données multimodales car les données des DME incluent à la fois des informations structurées, telles que les résultats de laboratoire, et des informations non structurées, comme les notes cliniques ; et (v) l'impact des facteurs externes (e.g., mode de vie, environnement) pouvant entraîner une variabilité et de l'incertitude dans les prédictions. Relever ces défis est essentiel pour développer des systèmes de prédiction des trajectoires des patients à la fois précis et fiables, capables d'assister les médecins dans la prise de décision en fournissant des prévisions complètes basées sur l'historique clinique d'un patient.

Dans cette optique, nous nous concentrons sur l'amélioration de la précision des systèmes de pronostic médical automatisés, en particulier pour la prédiction des diagnostics futurs à partir des dossiers médicaux. Les systèmes de codage actuels, tels que la Classification Internationale des Maladies (CIM ou ICD en anglais) [2], ne capturent souvent pas pleinement la richesse des informations contenues dans les notes cliniques, ce qui peut entraîner une perte d'informations précieuses pour la prédiction des trajectoires des patients. Pour surmonter ce problème, nous proposons une approche visant à améliorer la précision des prévisions de codes de diagnostic en intégrant les embeddings des notes cliniques dans les Transformers, qui reposent habituellement uniquement sur les codes médicaux. Cette méthode intègre un facteur discriminant qui réduit les erreurs de prédiction en enrichissant la représentation des embeddings. Cela permet également de récupérer des informations précieuses souvent perdues dans des systèmes de codification tels que ICD. En intégrant un contexte supplémentaire, notre approche répond aux défis liés à la compréhension des raisons derrière les prescriptions de médicaments, les procédures effectuées et les diagnostics posés.

Cet article est organisé comme suit : la Section 2 passe en revue la littérature. La Section 3 décrit notre approche, y compris le processus de génération des embeddings et leur intégration dans les Transformers. Dans la Section 4, nous présentons nos résultats expérimentaux. Enfin, la Section 5 conclut cet article et présente les travaux futurs.

2 Etat de l'art

Diverses méthodes, qu'elles soient basées sur l'apprentissage profond ou des approches traditionnelles, ont été explorées pour prédire les trajectoires des patients. Parmi elles, *Doctor AI* (Choi et al., 2016a), un modèle temporel basé sur des réseaux de neurones récurrents (RNN), développé et appliqué à des données EHR (*Electronic Health Records*) longitudinales horodatées. *Doctor AI* prédit les codes médicaux d'un patient et estime le temps jusqu'à la prochaine visite. Cependant, il est limité par une largeur de fenêtre fixe, ce qui s'avère inadapté, car le futur diagnostic d'un patient peut dépendre de conditions médicales situées en dehors de cette fenêtre. *LIG-Doctor* (Rodrigues-Jr et al., 2021), une architecture de réseau

1. https://www.cdc.gov/nchs/icd/icd-10-cm/index.html
2. https://www.who.int/standards/classifications/classification-of-diseases

neuronal artificiel conçue pour prédire efficacement les trajectoires de patients en utilisant des réseaux récurrents minimaux bidirectionnels *MGRU*. *MGRU* traitent la granularité des codes ICD-9, mais souffrent des mêmes limties que *Doctor AI*. *RETAIN* (Choi et al., 2016b) est un modèle interprétable grâce à un mécanisme d'attention inversée, tandis que *DeepCare* (Pham et al., 2017) utilise des LSTM (Long Short-Term Memory) pour la prédiction et la gestion des risques médicaux. *Deep Patient* (Miotto et al., 2016) utilise des Autoencodeurs non supervisés, bien qu'il ne considère pas la temporalité des trajectoires. En parallèle, des méthodes plus classiques comme les chaînes de Markov (Severson et al., 2020), les réseaux bayésiens (Longato et al., 2022), et les processus de Hawkes (Lima, 2023) ont été explorées, mais souffrent de complexité computationnelle face à des données massives. L'introduction des modèles Transformers a marqué une avancée, avec *Clinical GAN* (Shankar et al., 2023) un modèle GAN basé sur cette architecture pour contrer le biais d'exposition (Arora et al., 2022). Cependant, ces modèles souffrent de problèmes d'évolutivité tels que l'instabilité de l'entraînement et la non-convergence (Saad et al., 2024).

Malgré le développement de diverses approches pour la prédiction des codes médicaux, il est important de noter que la plupart des modèles proposés ont été entraînés sur des dossiers médicaux électroniques (DME) contenant uniquement des données structurées sur des diagnostics et des procédures, telles que les codes ICD et CCS. Cependant, ces données omettent certaines informations contextuelles essentielles, telles que les raisonnements médicaux et les nuances spécifiques aux patients, qui peuvent être capturées à travers les notes cliniques. De plus, la comparaison des résultats entre différentes études pose plusieurs défis, notamment la variation des ensembles de données, la taille des échantillons de test, le manque de standardisation, les différences de prétraitement et les incohérences dans le mappage des codes. Ces défis soulignent la nécessité d'une réflexion approfondie dans l'évaluation des résultats. Pour améliorer la comparabilité et la reproductibilité des recherches sur la prédiction des trajectoires des patients, il est crucial de standardiser les ensembles de données, les méthodes de prétraitement et les métriques d'évaluation.

3 Méthodologie proposée

Dans cette section, nous décrivons notre approche pour la prédiction des trajectoires des patients, qui s'appuie sur le jeu de données MIMIC-IV [3]. Nous détaillons notre méthodologie, y compris le prétraitement des données, l'architecture du modèle et l'intégration des notes cliniques.

3.1 Prétraitement des données

Le prétraitement des données MIMIC-IV comprend plusieurs opérations : (i) extraction des diagnostics, procédures et médicaments, (ii) sélection des patients avec au moins deux visites, (iii) exclusion des patients sans les trois types de codes médicaux (Shankar et al., 2023), (iv) utilisation du CCSR (Clinical Classification Software Refined) pour mapper les diagnostics ICD-10-CM en catégories cliniquement significatives, équilibrant les catégories CCS (Clinical Classifications Software) de l'ICD-9-CM avec la spécificité de l'ICD-10-CM, ainsi que les

3. https://physionet.org/content/mimiciv/2.1/

procédures ICD-10-PCS, en tirant parti de la spécificité et de la taxonomie du schéma de codage ICD-10-PCS, (v) suppression des codes peu fréquents (seuil de 5) (Edin et al., 2023), et (vi) ordonnancement temporel des événements pour créer des trajectoires séquentielles.
Le tableau 1 présente les statistiques des codes avant et après l'application des étapes de traitement. La figure 1 illustre le déséquilibre des données, avec un plus grand nombre de patients ayant effectué une seule visite comparé à ceux ayant plus de deux visites. Travailler avec des ensembles de données textuelles limités pose des défis, notamment à cause des tokenizers sous-mot qui fragmentent différemment des tokens similaires en raison de légères variations structurelles. Par conséquent, nous avons standardisé les notes cliniques en unifiant les abréviations médicales (par exemple, « hr », « hrs » et « hr(s) » en « heures »), en supprimant les accents, en convertissant les caractères danois (comme « æ » en « ae »), et en mettant toutes les notes en minuscules, suivant l'approche de (Alsentzer et al., 2019).

Phase	Statistiques	Procedure codes	Diagnosis codes	Drug codes
Lors du chargement	Distinct codes	8482	15763	1609
	Mean ± std / visit	3.03 ± 2.81	12.50 ± 7.67	24.12 ± 28.19
Après pré-traitement	Distinct codes	470	762	1609
	Mean ± std / visit	2.99 ± 2.77	13.18 ± 8.58	24.12 ± 28.19

TAB. 1 – *Statistiques des codes avant et après traitement*

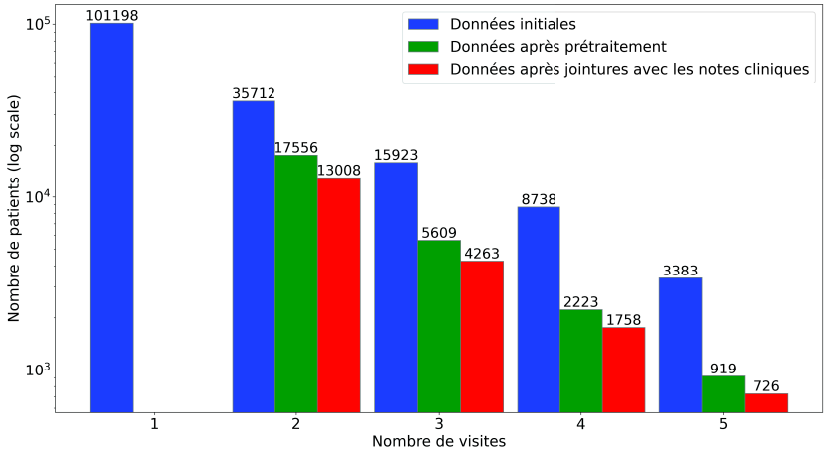

FIG. 1 – *Distribution des échantillons selon le nombre de visites par patient à chaque phase de traitement*

3.2 Intégration des notes cliniques

Nous proposons d'incorporer les embeddings des notes cliniques dans des modèles basés sur les modèles Transformers, qui se sont traditionnellement limités aux codes cliniques, à travers notre modèle, *Clinical Mosaic*. Cette approche vise à tirer parti des données structurées

et non structurées, offrant ainsi une vue plus complète de l'état clinique et de l'historique d'un patient. De plus, l'intégration de ces connaissances pourrait renforcer la capacité d'un modèle à comprendre et prédire les trajectoires des patients de manière plus précise.

3.2.1 Modèle Clinical Mosaic

Pour exploiter efficacement les informations contenues dans les notes cliniques, il est crucial d'obtenir des représentations sous forme de vecteurs. Les modèles BERT (Devlin et al., 2018), et plus particulièrement *Clinical BERT* (Alsentzer et al., 2019), ont prouvé leur capacité à capturer des représentations sémantiques pertinentes dans le domaine médical. *Clinical BERT* est un modèle pré-entraîné sur MIMIC-III, spécifiquement conçu pour les notes médicales. Cependant, ce modèle présente certaines limitations : il est entraîné sur des séquences de longueur limitée (128 tokens) et repose sur une version obsolète des données, à savoir MIMIC-III plutôt que MIMIC-IV.

Pour surmonter les limitations des modèles existants, nous proposons un modèle nommé **Clinical Mosaic**, s'appuyant sur l'architecture *MosaicBERT* (a Bidirectional Encoder Optimized for Fast Pretraining) (Portes et al., 2023) et pré-entraîné sur 331794 notes cliniques. Ce modèle étend la longueur des séquences à 512 tokens et est réentraîné sur la base de données MIMIC-IV-NOTES 2.2, plus récente et diversifiée. De plus, nous intégrons des améliorations récentes telles que *Attention with Linear Biases* (ALiBi) qui permet l'extrapolation à des séquences plus longues. Les paramètres d'entraînement sont résumés dans les tableaux 2 et 3.

Paramètre	**Valeur**
Effective Batch Size	224
Training Steps	80,000
Longueur de séquence	512 tokens
Probabilité de masquage	30%

TAB. 2 – *Paramètres généraux*

Détails de l'optimiseur	**Valeur**
Optimiseur	ADAMW
Learning Rate initial	$5e{-4}$
Learning Rate Scheduler	Linear warmup (33k steps) cosine annealing (46k steps)
Learning Rate final	$1e{-5}$

TAB. 3 – *Paramètres de l'optimiseur*

3.2.2 Évaluation du raisonnement clinique de Clinical Mosaic

Pour évaluer les performances de Clinical Mosaic, nous avons affiné le modèle sur le jeu de données *Medical Natural Language Inference* (MedNLI) (Romanov et Shivade, 2018). MedNLI est un jeu de données conçu pour les tâches d'inférence de langage naturel dans le domaine clinique, dérivé des notes cliniques de MIMIC-III. Il se compose de 14 049 paires de prémisses (i.e. une phrase extraite d'un rapport clinique, décrivant des observations ou résultats) et d'hypothèses (i.e. une déclaration à évaluer en fonction de la prémisse), avec pour objectif de classer la relation entre chaque paire comme étant une implication, une contradiction ou neutre. La tâche MedNLI évalue plusieurs aspects essentiels de la compréhension du langage clinique, notamment la compréhension sémantique de la terminologie médicale, le raisonnement logique dans un contexte clinique, ainsi que la capacité à discerner des relations nuancées entre des énoncés cliniques. Les performances sur ce jeu de données servent d'indicateur de la capacité d'un modèle à comprendre et à raisonner sur le langage clinique, une base

cruciale pour la prédiction des trajectoires des patients. Le tableau 4 présente les performances de *Clinical Mosaic* en termes de précision (*accuracy*) sur la tâche MedNLI en comparaison avec d'autres modèles de pointe, y compris le modèle original *Clinical BERT* (Alsentzer et al., 2019).

Modèle	Précision
BERT	77.6%
BioBERT	80.8%
Discharge Summary BERT	80.6%
Clinical Discharge BERT	84.1%
Bio+Clinical BERT	82.7%
Clinical Mosaic	**86.5%**

TAB. 4 – *Comparaison des performances des variantes de BERT et de Clinical Mosaic*

Les résultats montrent que Clinical Mosaic atteint une précision supérieure (86,5%) par rapport aux modèles existants, y compris le Clinical BERT original (84,1%). Cette amélioration suggère que nos optimisations du modèle et notre approche de pré-entraînement ont renforcé ses capacités de compréhension du langage clinique.

3.3 Fusion des Représentations Cliniques

Pour évaluer l'impact de l'intégration des embeddings des notes cliniques dans un modèle Transformer de type Encodeur-Décodeur, nous avons mené une série d'expériences afin d'identifier le point de fusion optimal au sein de l'architecture. Nous avons opté pour l'introduction des embeddings des notes cliniques dès la première couche du modèle, comme illustré dans la figure 2, avant l'application des mécanismes d'attention. Cette approche permet d'exploiter pleinement les capacités des couches d'attention multi-têtes, favorisant une meilleure interaction et une représentation plus efficace des embeddings fusionnés.

Lors de la génération des embeddings des notes cliniques, chaque couche des encodeurs BERT produit une représentation différente des séquences introduites. En nous appuyant sur le travail de (Devlin et al., 2018), qui a analysé l'impact de l'agrégation des embeddings issus des K dernières couches, nous avons déterminé que l'utilisation de 6 couches permettait d'obtenir les meilleurs résultats pour les tâches en aval, tout en restant efficace sur le plan computationnel. Par conséquent, dans nos expériences, nous avons fixé ce paramètre à 6 couches de représentation.

Pour la génération des embeddings, plusieurs stratégies peuvent être adoptées :

— **Moyenne sur les couches et les visites (MEAN)** : Cette approche calcule la moyenne des embeddings sur les 6 couches et toutes les visites, créant une représentation unifiée qui capture le contexte global des visites multiples, tout en lissant les bruits et en révélant les motifs communs.

— **Moyenne uniquement sur les couches (CONCAT)** : Les embeddings sont moyennés uniquement sur les couches, offrant des représentations multicouches pour chaque visite tout en réduisant la dimensionnalité de manière significative.

— **Utilisation d'une projection (Projection)** : Les embeddings des 6 couches de chaque visite sont projetés dans un espace de dimension inférieure via une couche linéaire

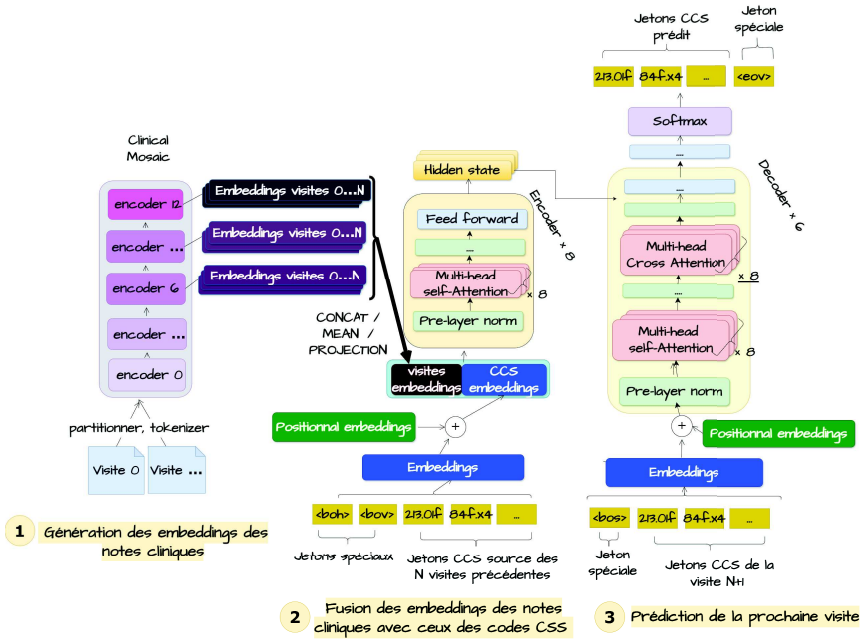

FIG. 2 – *Architecture pour intégrer les notes*

avec activation GeLU, suivie d'une autre couche linéaire. Cette méthode réduit la dimensionnalité tout en capturant les relations complexes entre les visites (figure 3).

Après avoir généré les embeddings à partir des stratégies précédentes, nous les intégrons avec les embeddings des codes CCS comme illustré dans la figure 2. Cette intégration utilise l'architecture Transformer, où les codes CCS peuvent prêter attention aux embeddings des notes cliniques via un mécanisme de self-attention, créant une représentation unifiée. Le décodeur du Transformer, à l'aide d'une cross-attention causale, utilise cette représentation pour prédire les diagnostics des visites futures. Cette approche permet au modèle de combiner efficacement les données structurées (codes CCS) et non structurées (notes cliniques), offrant une vue complète de l'historique clinique du patient et visant à améliorer les performances prédictives pour les trajectoires des patients.

4 Expérimentations

Dans cette section, nous présentons les expériences menées pour évaluer notre approche sur les jeux de données MIMIC-IV et MIMIC-IV-NOTES (37k échantillons après prétraitement), qui, bien que publiés séparément, partagent plusieurs colonnes permettant de les associer et ainsi obtenir à la fois les notes cliniques et les codes de visite d'un patient.

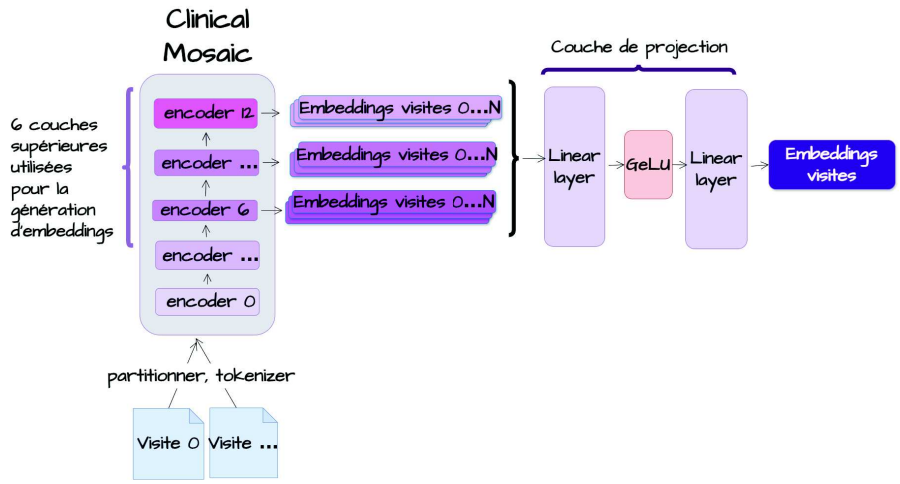

FIG. 3 – *Approche utilisant une couche de projection*

4.1 Métriques

Nous évaluons nos modèles à l'aide de la précision moyenne à K (MAP@K) et du rappel moyen à K (MAR@K) (équation 1) pour K = 20, 40, et 60. Ces métriques, particulièrement adaptées à notre problématique de recommandation où l'ordre des éléments est crucial, permettent une comparaison directe avec les travaux antérieurs, bien que certaines études se limitent à une seule métrique (Rodrigues-Jr et al., 2021).

$$\text{MAP@K} = \frac{1}{|Q|}\sum_{u=1}^{|Q|}\frac{1}{\min(m,K)}\sum_{k=1}^{K}P(k)\cdot rel(k), \quad \text{MAR@K} = \frac{1}{|Q|}\sum_{u=1}^{|Q|}\frac{1}{m}\sum_{k=1}^{K}rel(k) \tag{1}$$

Où $|Q|$ est le nombre de séquences cibles, m est le nombre d'éléments pertinents dans une séquence cible, K est le rang limite, $P(k)$ est la précision au rang k, et $rel(k)$ est une fonction valant 1 si l'élément au rang k est pertinent, 0 sinon.

4.2 Modèles comparés

Nous comparons notre approche au modèle Transformer sans intégration de notes cliniques (Vaswani et al., 2017) et aux modèles de l'état de l'art présentés dans le tableau 5. Les modèles ont été reproduits avec le framework Pytorch, en utilisant les codes et publications associés. Tous ont été évalués via une validation croisée à 5 folds et des intervalles de confiance à 95 %. Le code source est mis à disposition à des fins de reproductibilité [1].

4.3 Résultats

Les résultats obtenus sont présentés dans le tableau 6. La première observation est que l'injection des embeddings des notes cliniques dans l'architecture améliore significativement

Modèle	Configuration
Lig Doctor [1]	Dimension d'embeddings et cachée = 714 (taille de l'étiquette). Deux couches linéaires fusionnant le contexte bidirectionnel, suivies d'une couche softmax. Entraînement : 100 époques (convergence à 13), patience=10, batch=512, optimiseur Adadelta.
Doctor AI [2]	Dimension cachée et d'embedding = 2000, dropout=0,5. Entraînement : 20 époques, batch=384, optimiseur Adadelta.
Clinical GAN [3]	Générateur : encodeur-décodeur (3 couches, 8 têtes, dim=256). Discriminateur : encodeur Transformer (1 couche, 4 têtes). Entraînement : 100 époques (convergence à 11), batch=8, Adam (générateur), SGD (discriminateur), scheduler Noam.

[1](Choi et al., 2016a), [2](Rodrigues-Jr et al., 2021), [3](Shankar et al., 2023)

TAB. 5 – *Modèles de l'état de l'art*

modeles	K = 20		K = 40		K = 60	
	MAR	MAP	MAR	MAP	MAR	MAP
Projection	0.425(5)	0.556(21)	0.439(4)	0.556(21)	0.439(4)	0.556(21)
Concat	0.420(6)	0.569(6)	0.425(5)	0.571(6)	0.425(5)	0.571(6)
Mean	0.416(6)	0.538(84)	0.423(6)	0.567(17)	0.423(6)	0.567(17)
Clinical GAN[1]	0.410(5)	0.558(11)	0.414(5)	0.559(12)	0.414(5)	0.559(12)
Transformer	0.398(23)	0.565(23)	0.405(25)	0.566(23)	0.405(25)	0.566(23)
LIG-Doctor[2]	0.267(48)	0.474(94)	0.361(42)	0.431(87)	0.420(37)	0.402(80)
Doctor AI[3]	0.233(5)	0.206(46)	0.233(5)	0.207(47)	0.233(5)	0.207(47)

[1](Shankar et al., 2023), [2](Rodrigues-Jr et al., 2021), [3](Choi et al., 2016a)

Note : Les valeurs sont présentées sous la forme moyenne(écart-type). Par exemple, 0,425(5) représente 0,425 ± 0,005.

TAB. 6 – *Performances des différents modèles utilisant MAP@k et MAR@k. Les valeurs sont présentées sous la forme moyenne(écart-type à la dernière décimale).*

les performances, en particulier en termes de MAR@K. Cependant, nous considérons que cette amélioration pourrait être entravée par la taille restreinte de notre jeu de données (37k échantillons), ce qui empêche le modèle d'apprendre à exploiter pleinement les représentations des embeddings injectés. La stratégie de moyennage des couches d'embeddings et des visites (*Mean*) donne le MAR@K le plus faible parmi les approches d'injection d'embeddings. Cela peut être dû à la compression excessive de l'information, entraînant une perte d'informations. Toutefois, cette méthode est la plus efficace computationnellement, car elle ajoute seulement un vecteur. Cette efficacité est importante, notamment en raison de la complexité computationnelle en $O(N^2)$ du mécanisme d'attention des Transformers.

La méthode de concaténation, qui consiste à faire la moyenne uniquement des couches d'embeddings, apporte des améliorations significatives en termes de MAP@K et surpasse également toutes les méthodes de la littérature en termes de MAR@K, avec des scores de variance faibles sur les différentes divisions de validation croisée. Cela peut être justifié par deux raisons principales : d'une part, la représentation riche obtenue à partir des notes, moyennée sur plusieurs couches, renforce la richesse des informations tout en préservant les éléments critiques, contrairement à l'approche *Mean* qui perd de l'information. D'autre part, cette approche permet également au modèle d'être plus sélectif dans le traitement des informations, en exploitant les éléments indépendants des différentes visites médicales. *Lig Doctor* est conçu comme une

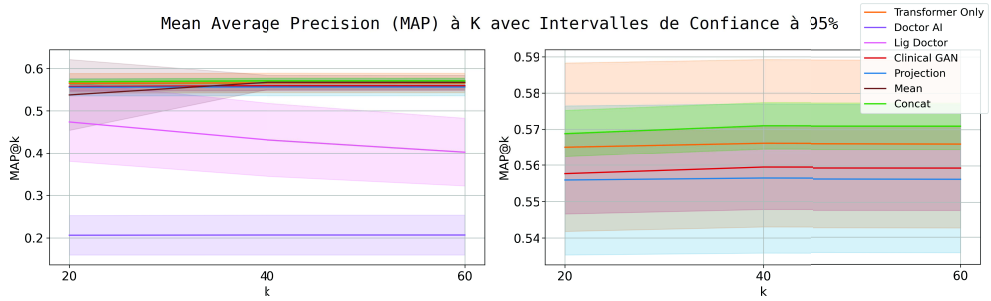

FIG. 4 – *Mean average precision @ 20, 40 et 60 pour différents modèles*

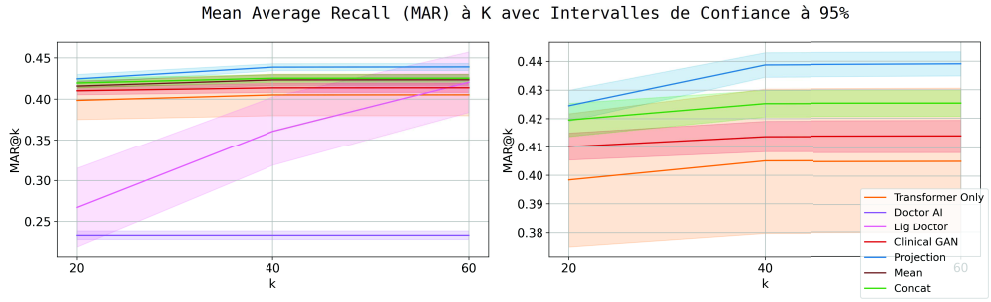

FIG. 5 – *Mean average recall @ 20, 40 et 60 pour les différents modèles*

tâche de classification, dans laquelle une couche linéaire est utilisée pour prédire les diagnostics suivants. Cette configuration introduit deux distinctions importantes dans l'évaluation. Premièrement, le modèle ne génère pas de prédictions dans un ordre spécifique, ce qui rend le calcul direct de métriques comme MAP@K impossible. Pour résoudre ce problème, nous proposons de trier les logits afin d'établir un ordre de génération. Toutefois, comme le modèle n'a pas été entraîné en tenant compte de cette information, les performances ne s'améliorent pas efficacement lorsque K augmente, comme le montre la figure 4. Deuxièmement, la classification empêche les prédictions répétitives, ce qui améliore les résultats en MAR@K. La figure 5 montre que *Lig Doctor* tire profit de l'augmentation de K.

Les résultats de *Doctor AI* sont inférieurs à ceux des autres modèles, car le modèle repose sur une seule couche GRU. La performance semble en effet affectée par l'augmentation de la dimension de l'espace de prédiction, et des améliorations pourraient être réalisées en augmentant les dimensions cachées ainsi que le nombre de couches GRU. *Clinical GAN* montre de bons résultats en termes de MAP@K, mais a des difficultés à générer un ensemble plus large de prédictions pertinentes, comme l'indiquent ses scores plus faibles de MAR@K. Ce modèle présente également une instabilité lors de l'entraînement, un problème fréquent des architectures basées sur des GANs, limitant leur mise à l'échelle. Cette limitation n'a pas pu être pleinement explorée en raison de la taille restreinte du jeu de données utilisé, laissant cette question ouverte pour de futures recherches.

5 Conclusion

Dans cet article, nous proposé une approche pour prédire les trajectoires de maladies des patients à partir de leurs dossiers médicaux électroniques (DME), qui intègre les embeddings des notes cliniques dans des modèles Transformers. Les résultats des expérimentations ont montré que l'approche proposée surpasse les modèles traditionnels basés uniquement sur les codes structurés en termes de précision. Pour les travaux futurs, nous prévoyons d'étendre notre approche à l'intégration de données multimodales et de relever les défis associés au traitement des données textuelles bruitées ou incomplètes, ainsi qu'à la gestion des prédictions non ordonnées.

Références

Alsentzer, E., J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, et M. McDermott (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*.

Arora, K., L. E. Asri, H. Bahuleyan, et J. C. K. Cheung (2022). Why exposure bias matters : An imitation learning perspective of error accumulation in language generation. *arXiv preprint arXiv :2204.01171*.

Choi, E., M. T. Bahadori, A. Schuetz, W. F. Stewart, et J. Sun (2016a). Doctor ai : Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318.

Choi, E., M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, et W. F. Stewart (2016b). RETAIN : an interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29 : Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3504–3512.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Edin, J., A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, et L. Maaløe (2023). Automated medical coding on mimic-iii and mimic-iv : A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2572–2582.

Egger, J., C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, et J. Kleesiek (2022). Medical deep learning—a systematic meta-review. *Computer methods and programs in biomedicine 221*, 106874.

Lima, R. (2023). Hawkes processes modeling, inference, and control : An overview. *SIAM Review 65*(2), 331–374.

Longato, E., M. L. Morieri, G. Sparacino, B. Di Camillo, A. Cattelan, S. L. Menzo, M. Trevenzoli, A. Vianello, G. Guarnieri, F. Lionello, et al. (2022). Time-series analysis of multidimensional clinical-laboratory data by dynamic bayesian networks reveals trajectories of covid-19 outcomes. *Computer Methods and Programs in Biomedicine 221*, 106873.

Miotto, R., L. Li, B. A. Kidd, et J. T. Dudley (2016). Deep patient : an unsupervised representation to predict the future of patients from the electronic health records. *Scientific*

reports 6(1), 1–10.

Pham, T., T. Tran, D. Phung, et S. Venkatesh (2017). Predicting healthcare trajectories from medical records : A deep learning approach. *Journal of biomedical informatics 69*, 218–229.

Portes, J., A. Trott, S. Havens, D. King, A. Venigalla, M. Nadeem, N. Sardana, D. Khudia, et J. Frankle (2023). Mosaicbert : A bidirectional encoder optimized for fast pretraining. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et S. Levine (Eds.), *Advances in Neural Information Processing Systems 36 : Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Rodrigues-Jr, J. F., M. A. Gutierrez, G. Spadon, B. Brandoli, et S. Amer-Yahia (2021). Lig-doctor : Efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks. *Information Sciences 545*, 813–827.

Romanov, A. et C. Shivade (2018). Lessons from natural language inference in the clinical domain. In E. Riloff, D. Chiang, J. Hockenmaier, et J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 1586–1596. Association for Computational Linguistics, doi: 10.18653/v1/D18-1187.

Saad, M. M., R. O'Reilly, et M. H. Rehmani (2024). A survey on training challenges in generative adversarial networks for biomedical image analysis. *Artificial Intelligence Review 57*(2), 19.

Severson, K. A., L. M. Chahine, L. Smolensky, K. Ng, J. Hu, et S. Ghosh (2020). Personalized input-output hidden markov models for disease progression modeling. In *Machine learning for healthcare conference*, pp. 309–330. PMLR.

Shankar, V., E. Yousefi, A. Manashty, D. Blair, et D. Teegapuram (2023). Clinical-gan : Trajectory forecasting of clinical events using transformer and generative adversarial networks. *Artificial Intelligence in Medicine 138*, 102507.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Summary

The prediction of patient disease trajectories using Electronic Health Records (EHRs) is challenging due to non-stationarity, the granularity of medical codes, and difficulties in integrating multimodal data. Current models often overlook critical insights in unstructured data, primarily relying on structured diagnosis codes. In this paper, we propose a novel approach that incorporates unstructured clinical notes into deep learning models for sequential disease prediction. By embedding clinical notes in Transformer-based models, we provide a richer representation of patient histories, enhancing accuracy in predicting future diagnoses. Our experiments show significant improvements in predictive performance compared to traditional models relying solely on structured codes.