Le texte clinique suffit-il ? Une étude multimodale de la prédiction de la mortalité dans l'insuffisance cardiaque

Oumaima El Khettari[1] Emmanuel Morin[1] Pierre Zweigenbaum[2]
(1) Nantes Univ., École Centrale Nantes, CNRS, LS2N, UMR 6004, 44000 Nantes, France
(2) Université Paris-Saclay, CNRS, LISN, Orsay, France
oumaima.el-khettari@univ-nantes.fr, emmanuel.morin@univ-nantes.fr, pz@lisn.fr

RÉSUMÉ

Nous évaluons des approches unimodales et multimodales pour la prédiction de mortalité à 3 mois dans l'insuffisance cardiaque sur une cohorte française. L'enrichissement des représentations CLS par des entités nommées améliore les performances textuelles. La fusion multimodale atteint les meilleurs résultats (F1=48,5%, AUC=83,8%), tandis que les grands modèles de langue restent peu fiables.

ABSTRACT

Is Clinical Text Enough ? A Multimodal Study on Mortality Prediction in Heart Failure Patients

We evaluate unimodal and multimodal approaches for 90-day mortality prediction in heart failure on a French cohort. Enriching CLS embeddings with named entity representations improves text-only performance. Multimodal fusion achieves the best results (F1=48.5%, AUC=83.8%), while large language models remain unreliable.

MOTS-CLÉS : TAL clinique, données structurées, fusion multimodale, insuffisance cardiaque.

KEYWORDS: Clinical NLP, structured data, multimodal fusion, heart failure.

ARTICLE ACCEPTÉ À : The Fifteenth biennial Language Resources and Evaluation Conference (LREC 2026).
URL : http://www.lrec-conf.org/proceedings/lrec2026/pdf/2026.lrec2026-1.14.pdf

1 Introduction

La prédiction de la mortalité à court terme dans l'insuffisance cardiaque (IC) demeure un enjeu clinique majeur. Les dossiers médicaux électroniques (DME) offrent des signaux pronostiques répartis entre deux modalités complémentaires : les variables structurées (biologie, démographie, traitements) et les notes cliniques libres, décrivant l'évolution du patient et le contexte médical. Pourtant, la plupart des travaux, souvent basés sur le jeu de données MIMIC en anglais (Johnson *et al.*, 2016; Nargesi *et al.*, 2025), privilégient des approches unimodales, soit centrées sur les variables structurées, soit sur des représentations globales du texte (Hashir & Sawhney, 2020; Memarzadeh *et al.*, 2022), en négligeant l'information à granularité fine portée par les entités cliniques. Par ailleurs, les DME

français présentent des spécificités terminologiques et structurelles encore peu explorées (D'hondt *et al.*, 2015; Gaschi *et al.*, 2023).

Dans ce travail, nous proposons une évaluation systématique sur une cohorte française de patients IC, en comparant des approches texte seul, données structurées seules et multimodales, ainsi que des grands modèles de langue (GML). Nous intégrons explicitement des représentations au niveau des entités extraites des notes cliniques, afin d'évaluer leur contribution à la prédiction de la mortalité à court terme.

2 Données et Ressources

Le jeu de données comprend 2 254 séjours hospitaliers de patients atteints d'insuffisance cardiaque (IC), extraits du service de cardiologie de l'Hôpital Paris Saint-Joseph. Chaque séjour est associé à une étiquette binaire indiquant un décès dans les trois mois suivant la sortie (11% de cas positifs), reflétant un fort déséquilibre de classes. Les notes cliniques, rédigées en français dans le système DxCare, constituent la modalité textuelle.

Les **données structurées** regroupent 115 variables collectées à l'admission (démographie, comorbidités, biologie, traitements), réduites à 41 variables après sélection par LASSO (Muthukrishnan & Rohini, 2016). La cohorte est caractérisée par un âge médian de 81 ans, représentatif d'une population âgée atteinte d'IC.

Les **entités nommées** (EN) sont extraites des notes cliniques selon un protocole en deux étapes. Une pré-annotation heuristique, fondée sur des ressources terminologiques (UMLS, BDPM), des expressions régulières et des patrons morphosyntaxiques (Barthet *et al.*, 2023), est d'abord appliquée. Elle est ensuite complétée par un modèle de reconnaissance d'entités nommé (BERT) entraîné sur 23 types d'entités cliniques (e.g., Pathologie, Traitement, Anatomie, Signe/Symptôme). Ce modèle atteint un F1 micro de 0,78 sur un sous-ensemble annoté par un expert, garantissant une qualité d'annotation satisfaisante.

3 Méthodes

Texte seul. Nous évaluons des représentations textuelles issues de modèles biomédicaux français (CamemBERT-bio (Touchent *et al.*, 2023), DrBERT (Labrak *et al.*, 2023)). Le vecteur `[CLS]` sert de représentation globale du document. Les entités nommées sont encodées par la moyenne des vecteurs de leurs sous-mots, puis agrégées par type. Nous comparons plusieurs stratégies de fusion entre CLS et entités : moyenne, somme, concaténation, pondération apprise, et fusion avec mécanisme de porte, permettant de moduler dynamiquement la contribution des entités.

Données structurées seules. Nous utilisons une régression logistique sur les 41 variables sélectionnées, après normalisation et imputation des valeurs manquantes. L'entraînement est réalisé avec pondération des classes et validation croisée stratifiée à 5 plis.

Multimodal. Les approches multimodales combinent représentations textuelles (CLS seul ou enrichi par entités) et variables structurées. Nous évaluons des stratégies simples (concaténation directe, fusion tardive par moyenne ou empilement) ainsi que des mécanismes appris : fusion avec mécanisme de

porte, double attention croisée et variantes bidirectionnelles, permettant de modéliser les interactions entre modalités.

GLMs. Nous évaluons Mistral-7B-Instruct (Jiang *et al.*, 2024), Qwen2.5-7B-Instruct (Qwen *et al.*, 2025) et MedGemma-4B-Instruct (Sellergren *et al.*, 2025) sans démonstrations, via des instructions en français contenant les notes cliniques, les variables structurées (brutes ou textualisées), ou leur combinaison. Les modèles génèrent directement une classe (0/1). Les performances sont évaluées par précision, rappel et F1 sur la classe positive, et par l'AUC-ROC pour les modèles supervisés.

4 Résultats et Discussion

Modèle	P	R	F1	AUC
Texte seul (CamemBERT-bio)				
CLS seul	28,9	48,8	36,2	75,1
Fusion par porte (CLS+EN)	37,1	42,3	**39,5**	75,4
Moy. pondérée (CLS+EN)	32,8	47,6	38,7	**75,8**
Données structurées seules				
Régr. logistique (41 var.)	26,2	69,4	38,0	**79,0**
Multimodal (CamemBERT-bio)				
Fusion par porte (CLS+EN+Struct.)	49,0	48,4	**48,5**	81,5
Fusion tardive (empilement)	62,8	29,1	39,5	**83,8**
GLMs (texte seul)				
MedGemma-4B	41,4	19,4	26,4	—
Mistral-7B	15,9	67,2	25,7	—

TABLE 1 – Résultats comparatifs. P/R/F1 sont calculés sur la classe positive (décès à 3 mois), AUC pour les modèles supervisés. Nous reportons uniquement les meilleures configurations pour chaque type de modèle (texte, structuré, multimodal). Les résultats sont présentés avec CamemBERT-bio, qui surpasse DrBERT dans nos expériences. EN = entités nommées.

Texte seul. Les représentations textuelles atteignent des AUC entre 70 et 76%. L'intégration des entités nommées améliore le F1 par rapport au CLS seul (+3,3 pts pour la fusion par porte), confirmant l'apport des représentations à grain fin.

Données structurées. La régression logistique sur 41 variables atteint une AUC de 79%, supérieure aux approches textuelles. Les prédicteurs les plus associés à la mortalité (âge, urée, CRP, troponine) et les variables protectrices (bêtabloquants, IEC/ARA II) sont cliniquement cohérents.

Multimodal. La fusion des deux modalités améliore systématiquement les performances. La fusion par porte avec entités atteint le meilleur F1 (48,5%), et la fusion tardive par empilement le meilleur AUC (83,8%). Des tests de significativité appariés ($p < 0{,}05$) confirment que la fusion multimodale améliore significativement la discrimination par rapport aux approches unimodales.

GLMs. Les performances restent faibles (F1 : 19–26%) et instables selon la modalité. Le texte clinique est mieux exploité que les variables structurées. L'adhérence aux instructions varie : Mistral génère des listes avec les données structurées, Qwen ajoute des explications, tandis que MedGemma

respecte le format. La textualisation améliore Mistral ($0 \rightarrow 19{,}4$) sans effet notable pour les autres.

5 Conclusion

La fusion multimodale supervisée enrichie d'entités nommées constitue l'approche la plus performante pour la prédiction de mortalité à 90 jours dans l'IC sur une cohorte française. Le texte clinique seul reste insuffisant, et les GLMs testés sont peu fiables pour cette tâche quel que soit le format d'entrée. Ces résultats soulignent l'intérêt des représentations à grain fin et des mécanismes de fusion appris pour intégrer données textuelles et structurées en milieu hospitalier francophone.

Remerciements

Ce travail a été soutenu par l'Agence Nationale pour la Recherche (ANR) dans le cadre du projet PREDHIC (ANR-21-CE23-0039).

Références

BARTHET V., AROULANDA M.-J., MONCEAUX-CACHARD L., JACQUIN C., GROUIN C., GUTTON J., HOCQUET G., DE GROOTE P., KOMAJDA M., MORIN E. & ZWEIGENBAUM P. (2023). La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation. In F. BOUDIN, B. DAILLE, R. DUFOUR, O. EL, M. HOUBRE, L. JOURDAN & N. KOOLI, Éds., *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, p. 1–7, Paris, France : ATALA.

D'HONDT E., TANNIER X. & NÉVÉOL A. (2015). Redundancy in French electronic health records : A preliminary study. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, p. 21–30.

GASCHI F., FONTAINE X., RASTIN P. & TOUSSAINT Y. (2023). Multilingual clinical NER : Translation or cross-lingual transfer ? In T. NAUMANN, A. BEN ABACHA, S. BETHARD, K. ROBERTS & A. RUMSHISKY, Éds., *Proceedings of the 5th Clinical Natural Language Processing Workshop*, p. 289–311, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.clinicalnlp-1.34.

HASHIR M. & SAWHNEY R. (2020). Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics*, **108**, 103489. DOI : https ://doi.org/10.1016/j.jbi.2020.103489.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). Drbert : Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique. In *18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 109–120 : ATALA.

MEMARZADEH H., GHADIRI N. & SHAHREZA M. L. (2022). Assessing mortality prediction through different representation models based on concepts extracted from clinical notes.

MUTHUKRISHNAN R. & ROHINI R. (2016). Lasso : A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)*, p. 18–20 : Ieee.

NARGESI A. A., ADEJUMO P., DHINGRA L. S., ROSAND B., HENGARTNER A., COPPI A., BENIGERI S., SEN S., AHMAD T., NADKARNI G. N. *et al.* (2025). Automated identification of heart failure with reduced ejection fraction using deep learning-based natural language processing. *Heart Failure*, **13**(1), 75–87.

QWEN, :, YANG A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., LI C., LIU D., HUANG F., WEI H., LIN H., YANG J., TU J., ZHANG J., YANG J., YANG J., ZHOU J., LIN J., DANG K., LU K., BAO K., YANG K., YU L., LI M., XUE M., ZHANG P., ZHU Q., MEN R., LIN R., LI T., TANG T., XIA T., REN X., REN X., FAN Y., SU Y., ZHANG Y., WAN Y., LIU Y., CUI Z., ZHANG Z. & QIU Z. (2025). Qwen2.5 technical report.

SELLERGREN A., KAZEMZADEH S., JAROENSRI T., KIRALY A., TRAVERSE M., KOHLBERGER T., XU S., JAMIL F., HUGHES C., LAU C. *et al.* (2025). Medgemma technical report. *arXiv preprint arXiv :2507.05201*.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). Camembert-bio : Un modèle de langue français savoureux et meilleur pour la santé. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux–articles longs*, p. 323–334 : ATALA.