



e-Algerian Journal
Health & Research

Disponible en ligne sur : ASJP

e-AJHR

<https://asjp.cerist.dz/en/PresentationRevue/968>



Article original

IA multimodale pour l'évaluation et la prédiction de la sévérité de la pancréatite aiguë.

Multimodal AI for assessing and predicting the severity of acute pancreatitis.

R.Madani^(a), M.Taieb^(b,d), N.Nait Slimane^(b,d), L.Sliman^(d), H.Dellys^(a), S.Artabaz^(a),
H.Mokeddem^(a)

^a Ecole Nationale Supérieure d'Informatique (Ex. INI) - ESI (Alger)

^b Service universitaire de chirurgie générale - EPH Ain Taya (Alger)

^c Grande école généraliste et numérique - EFREI (Paris)

^d Université des sciences de la santé Dr Youcef El Khatib, Département de médecine , Alger

Auteur correspondant **RAFIK MADANI**

Adresse e-mail : kr_madani@esi.dz

Résumé

Introduction : La pancréatite aiguë (PA) présente un défi clinique majeur en raison de son évolution imprévisible vers des formes sévères (PAS) associées à une mortalité élevée. Les systèmes de scores traditionnels (Ranson, BISAP, APACHE-II) manquent de sensibilité précoce et reposent sur des données unimodales. L'intelligence artificielle multimodale, intégrant données clinico-biologiques et d'imagerie, offre une opportunité d'améliorer la prédiction précoce de la sévérité. Cette étude vise à évaluer si un modèle d'IA fusionnant données biocliniques et tomodensitométries (TDM) peut surpasser significativement un modèle bioclinique seul pour la prédiction de la PAS.

Matériels et méthodes : Cohorte monocentrique de 255 patients admis pour PA (42 cas sévères, 16,5%). Un modèle TDM basé sur ResNet-50 a extrait des caractéristiques de 50 coupes axiales standardisées englobant les régions pancréatiques et péripancréatique. L'entraînement à double tâche (perte BCE + contrastive supervisée) a généré trois variants optimisés (Alpha: AUC, Beta: AUPRC, Gamma: Recall@P50) via validation croisée 5-plis. Un modèle bioclinique XGBoost "plug-and-play" (sans optimisation des hyperparamètres, n_estimators=100) a traité les données d'admission (24h) après exclusion des scores établis. La fusion multimodale a intégré les scores TDM comme caractéristiques additionnelles. Évaluation par métriques OOF agrégées et test de DeLong (IC 95%).

Résultats : Le modèle bioclinique baseline a atteint une AUROC OOF de 0,8196 (IC 95%: 0,7503-0,8888) avec excellente rétention per-fold/OOF (97-100%), comparable aux scores établis. Parmi les modèles TDM, V2-Alpha a démontré la meilleure généralisation OOF (0,7748 AUC, rétention 94,6%) et a surpassé Beta et Gamma dans les trois métriques. La fusion multimodale Alpha (MM-Alpha) a significativement amélioré les performances : AUROC 0,8492 (Δ AUC +0,0296, $p=0,0435$, IC 95%: 0,7866-0,9118), AUPRC 0,6398 (+14,9%), et Recall@P50 0,6190 (+52,9%, correspondant à 9 patients SAP supplémentaires détectés). MM-Alpha a également amélioré la rétention OOF du Recall@P50 de 71,8% à 90,6%. Les modèles MM-Beta et MM-Gamma n'ont montré aucune amélioration significative ($p=0,9445$ et $p=0,3299$), reflétant leur faible généralisation OOF unimodale (rétention 53-71%).

Conclusions : L'intégration de caractéristiques TDM extraites par apprentissage profond améliore significativement la prédiction de la sévérité de la PA, avec une augmentation remarquable de 53% du rappel à précision fixe. L'optimisation pour la discrimination globale (AUC) produit des signaux plus robustes et immédiatement exploitables en fusion multimodale que les optimisations métrique-spécifiques dans un contexte de cohorte limitée. Ces résultats suggèrent un potentiel clinique pour l'identification précoce des patients à haut risque, nécessitant validation externe multicentrique

© 2025e-AJHR. Tous droits réservés

Mots clés : Pancréatite aiguë, Intelligence artificielle, Apprentissage profond, Fusion multimodale, Tomodensitométrie, Prédiction de sévérité, XGBoost, ResNet-50.

Abstract

Background : Acute pancreatitis (AP) presents a major clinical challenge due to its unpredictable progression toward severe forms (SAP) associated with high mortality. Traditional scoring systems (Ranson, BISAP, APACHE-II) lack early sensitivity and rely on unimodal data. Multimodal artificial intelligence integrating clinical-biological and imaging data offers an opportunity to improve early severity prediction. This study aims to evaluate whether an AI model fusing bioclinical and computed tomography (CT) data can significantly outperform a bioclinical-only model for SAP prediction.

Materials and methods : Single-center cohort of 255 patients admitted for AP (42 severe cases, 16.5%). A ResNet-50-based CT model extracted features from 50 standardized axial slices encompassing pancreatic and peripancreatic regions. Dual-task training (BCE + supervised contrastive loss) generated three optimized variants (Alpha: AUC, Beta: AUPRC, Gamma: Recall@P50) via 5-fold cross-validation. A "plug-and-play" XGBoost bioclinical model (no hyperparameter optimization, n_estimators=100) processed admission data (24h) after excluding established scores. Multimodal fusion integrated CT scores as additional features. Evaluation used aggregated OOF metrics and DeLong test (95% CI).

Results: The bioclinical baseline achieved an OOF AUROC of 0.8196 (95% CI: 0.7503-0.8888) with excellent per-fold/OOF retention (97-100%), comparable to established scores. Among CT models, V2-Alpha demonstrated the best OOF generalization (0.7748 AUC, 94.6% retention) and outperformed Beta and Gamma across all three metrics. Multimodal Alpha fusion (MM-Alpha) significantly improved performance: AUROC 0.8492 (Δ AUC +0.0296, $p=0.0435$, 95% CI: 0.7866-0.9118), AUPRC 0.6398 (+14.9%), and Recall@P50 0.6190 (+52.9%, corresponding to 9 additional SAP patients detected). MM-Alpha also improved OOF retention for Recall@P50 from 71.8% to 90.6%. MM-Beta and MM-Gamma models showed no significant improvement ($p=0.9445$ and $p=0.3299$), reflecting their poor unimodal OOF generalization (53-71% retention).

Conclusion: Integration of deep learning-extracted CT features significantly improves AP severity prediction, with a remarkable 53% increase in recall at fixed precision. Optimization for global discrimination (AUC) produces more robust and immediately exploitable signals in multimodal fusion than metric-specific optimizations in limited-cohort settings. These results suggest clinical potential for early identification of high-risk patients, requiring multicenter external validation.

© 2025 e-AJHR. All rights reserved.

keywords: Acute pancreatitis, Artificial intelligence, Deep learning, Multimodal fusion, Computed tomography, Severity prediction, XGBoost, ResNet-50.

Introduction

La pancréatite aiguë (PA) est une affection inflammatoire aiguë du pancréas qui peut évoluer d'un épisode léger et spontanément résolutif à une maladie grave et potentiellement mortelle [2]. La forme sévère (PAS), est associée à une défaillance multiviscérale, une nécrose pancréatique et un taux de mortalité élevé. La prédiction précoce de la sévérité des PA est essentielle pour une prise en charge optimale des patients, l'admission en soins intensifs et une intervention thérapeutique rapide.

L'évaluation classique de la gravité reposait sur des systèmes de scores cliniques tels que le Ranson [1], BISAP [4], APACHE II [5] et MCTSI [6] associé à des marqueurs biologiques et d'imagerie. Cependant, ces outils manquent souvent de sensibilité dans la phase précoce de la maladie et sont limités par leur dépendance à des informations statiques ou unimodales. Parallèlement, les avancées en intelligence artificielle (IA) et en apprentissage automatique ont démontré des résultats

prometteurs dans la prédiction et la classification des PA [7]. En revanche, la plupart des modèles restent limités aux modalités cliniques ou d'imagerie, ne parvenant pas à saisir l'interaction complexe entre les facteurs systémiques et morphologiques qui déterminent l'évolution de la maladie.

Dans ce contexte, les modèles d'IA multimodaux qui intègrent des sources de données hétérogènes ont le potentiel de fournir une évaluation plus complète de l'état du patient [8]. En analysant conjointement les informations cliniques, biologiques et radiologiques, de tels systèmes peuvent modéliser les aspects fonctionnels et structurels de la pancréatite aiguë, permettant une prédiction plus précise et personnalisée de la gravité de la maladie.

La présente étude vise à explorer le potentiel de la multimodalité dans la prédiction de l'évolution et de la gravité de la PA. Nous avons développé un modèle d'IA multimodal combinant des données clinico-biologiques et des tomodensitogrammes (TDM). Le modèle TDM traite des coupes 2D extraites d'une région d'intérêt anatomique définie par nos soins englobant les zones pancréatiques et péripancréatiques intéressées par le processus pathologique. Un réseau ResNet-50 pré-entraîné sur des données TDM médicales a été utilisé pour l'extraction de caractéristiques, suivi d'une projection d'intégration et d'une moyenne de regroupement. Les sorties du modèle TDM ont ensuite été intégrées dans un modèle bioclinique basé sur XGBoost, afin d'évaluer la contribution des représentations basées sur l'imagerie à la performance prédictive globale.

Dans notre travail, nous adoptons une approche "prête à l'emploi" (plug-and-play) de la multimodalité dans la prédiction de la sévérité des PA. La sortie du modèle TDM est ajoutée comme simple variable supplémentaire à un modèle XGBoost standardisé, qui n'est pas optimisé par une recherche des hyperparamètres (les paramètres de configuration qui contrôlent comment le modèle apprend, tels que la vitesse d'apprentissage ou la complexité autorisée), et ce, en vue des limites sur la taille de la cohorte des patients. Ce banc d'essai XGBoost reste strictement identique, qu'il soit entraîné sur les données biocliniques seules ou sur les données multimodales.

Notre étude examine plusieurs configurations de fusion, désignées par alpha, beta et gamma, correspondant à différents critères de sélection du modèle TDM (ROC-AUC, AUPRC et rappel à une précision de 50%). Les modèles ont été comparés à l'aide du test de DeLong [9] avec des intervalles de confiance à 95 % pour évaluer la signification statistique.

L'objectif de ce travail est d'évaluer la valeur ajoutée brute de l'information TDM à l'information bioclinique. En utilisant ce modèle fixe, nous testons le signal morphologique pour un bénéfice immédiat et robuste. L'exploration du potentiel maximal par optimisation est réservée à de futurs travaux sur des cohortes plus larges.

Matériels et Méthodes

Population d'étude

L'ensemble des données utilisées pour cette étude se compose d'une cohorte de 255

patients admis pour PA, constituant un sous-ensemble de la cohorte initiale décrite par [1]. Pour chaque patient, les paramètres d'évaluation pris en compte sont :

1. **Données biocliniques** enregistrés à l'admission (24h) incluent : sexe, age, tabac, Hypertension, Diabète, SBP, DBP, MAP, Plaquettes, Globules blancs, Neutrophiles, Lymphocytes, NLR, Hématocrite, RDW, Lymphocytes, PDW, Créatinine, Bilirubine totale, Bilirubine directe, DTR, Urée, HDL, LDH, Calcium, Triglycérides, Glycémie, TYG, BUN, ALT, AST, γ -GT, ALP, PAB, Albumine, Potassium, A/G, PT, INR, APTT, CRP, CPR, CAR, RCR, MCTSI, Ranson, BISAP, SIRS, EP, SABP.
2. **Tomodensitométrie (TDM) 72 h**: TDM abdominales acquises pendant l'hospitalisation.

La prédiction de la sévérité est essentiellement une classification binaire de l'évolution de la maladie. Notre effectif scindé en deux groupes: **42 patients (16,5 %)** classés comme PAS (Classe 1) et **213 patients (83,5 %)** classés comme non-PAS (Classe 0) .

Dans ce qui suit, nous présenterons les éléments de notre solution en suivant la logique du pipeline que nous avons établi, , notre pipeline traite d'abord les images CT avec un modèle dédié pour obtenir une prédiction de sévérité. En parallèle, un modèle bio clinique est entraîné sur les données tabulaires. La prédiction CT est ensuite fusionnée avec celle du modèle bio clinique pour former le modèle multimodal. Un modèle bio clinique seul sert de référence pour la comparaison.

Partie 1 : Le modèle d'imagerie TDM

Traitement d'image et extraction de caractéristiques

Pour assurer une extraction cohérente des caractéristiques, une **région d'intérêt anatomique (ROI)** standardisée a été définie pour chaque patient, centrée autour du pancréas et incluant les zones péripancréatiques clés. Cette ROI était ancrée par des repères anatomiques :

- **Repère supra-pancréatique** : première coupe axiale passant par le bord supérieur du foie.
- **Région pancréatique** : toutes les coupes axiales contiguës incluant le pancréas.
- **Repère infra-pancréatique** : 10 coupes axiales contiguës sous pancréatiques..

Pour créer un volume d'entrée standardisé pour le modèle, un volume de taille fixe de **50 coupes 2D** a été extrait pour chaque patient. Ce bloc de 50 coupes a été positionné pour s'assurer qu'il **contenait entièrement la ROI spécifique au patient** telle que définie par les repères.

En pratique, cela signifiait que le volume de 50 coupes s'étendait souvent au-dessus du repère supra-pancréatique (la première coupe hépatique) et en dessous du repère infra-pancréatique (la 10e coupe sous le pancréas) pour satisfaire l'exigence des 50 coupes.

Chacune de ces 50 coupes a été passée à travers un modèle **ResNet-50 (modèle de traitement d'images)** [10], précédemment pré-entraîné sur un vaste jeu de données d'imagerie médicale dans une étude antérieure [11],

pour agir comme extracteur de caractéristiques. Le backbone (correspond à la partie centrale du réseau chargée d'extraire automatiquement les informations essentielles des images) du ResNet-50 a été figé jusqu'à la couche 4, ce qui signifie que seule la couche 4 et les couches entièrement connectées suivantes ont été affinées pour notre tâche spécifique de prédiction de sévérité. Les 50 embeddings (la représentation numérique compacte de l'image produite par le réseau, qui résume les informations essentielles pour la prédiction.) résultants, un pour chaque coupe ont ensuite été transformés via une tête de projection. Une opération de **pooling moyen** (une opération qui résume toutes les informations de l'ensemble des caractéristiques des 50 coupes en calculant leur moyenne, pour obtenir une représentation unique et compacte.) a été appliquée à ces 50 embeddings pour produire une seule représentation vectorielle de taille fixe pour l'ensemble du TDM du patient.

Prétraitement et augmentation d'image

Toutes les coupes 2D extraites ont d'abord été redimensionnées à 256x256 pixels. Il a été implémenter un pipeline d'augmentation de données pendant l'entraînement afin d'améliorer la robustesse et l'invariance du modèle. Pour les données d'entraînement, une découpe aléatoire de 224x224 a été effectuée. Cela a été suivi d'une série d'augmentations géométriques, notamment : des retournements horizontaux, des transformations affines mineures (rotation, échelle, translation), des déformations élastiques (appliquées avec une probabilité de 0,2), et du bruit gaussien (pour simuler la variabilité du scanner).

Ces augmentations améliorent l'invariance à l'orientation du patient et aux légers changements anatomiques. Les augmentations basées sur l'intensité (par exemple, luminosité/contraste) ont été délibérément évitées pour préserver la cohérence des valeurs d'unité Hounsfield (HU). Pour les données de validation, seul un prétraitement déterministe a été appliqué : une découpe centrale de 224x224 a été effectuée. Enfin, toutes les images (entraînement et validation) ont été normalisées min-max et converties en tenseurs avant d'être introduites dans le ResNet-50.

Entraînement à double tâche et sélection de modèle

Le modèle TDM, construit sur les caractéristiques poolées du ResNet-50, a été entraîné à l'aide d'une fonction d'objectif à double tâche pour apprendre des représentations robustes au niveau du patient. Pour faire face au déséquilibre significatif des classes dans l'ensemble de données (16,1 % de PAS), un échantillonneur de lots équilibré a été utilisé pendant l'entraînement. Cela garantit que chaque lot d'entraînement présente une distribution plus équilibrée des cas positifs (PAS) et négatifs (non-PAS), empêchant le modèle de devenir biaisé vers la classe majoritaire. La perte totale était une combinaison pondérée de : perte d'entropie croisée binaire (BCE), une perte de classification standard pour la tâche principale de prédiction de la gravité de la PA, perte contrastive supervisée (SupCon) [12]: cette perte a été appliquée aux embeddings du modèle pour encourager un espace de représentation plus discriminatif. Elle fonctionne en rapprochant les embeddings des patients de la même classe

de gravité (par exemple, tous les « Sévères »), tout en les éloignant simultanément des embeddings de la classe opposée.

Pour générer des prédictions fiables à utiliser dans l'étape de fusion ultérieure et éviter toute fuite d'information, nous avons appliqué une validation croisée à 5 plis (folds) sur l'ensemble d'entraînement. Les données ont été divisées en 5 groupes (appelés plis) : à chaque itération, le modèle était entraîné sur 4 plis et testé sur le 5^e. Pendant chaque entraînement, nous surveillions simultanément trois indicateurs de performance principaux, et chaque fois qu'un nouveau meilleur score était atteint pour l'un d'eux, le point de contrôle correspondant était enregistré.

À la fin des 5 plis, ce processus a produit trois ensembles distincts de prédictions hors échantillon (out-of-fold, OOF) couvrant l'ensemble des patients de l'entraînement, chaque ensemble regroupant les prédictions qui maximisent la métrique correspondante:

- **Modèle Alpha** : prédictions OOF issues des points de contrôle atteignant la meilleure aire sous la courbe ROC (AUROC)
 - **Modèle Beta** : prédictions OOF issues des points de contrôle atteignant la meilleure aire sous la courbe précision-rappel (AUPRC)
 - **Modèle Gamma** : prédictions OOF issues des points de contrôle atteignant le meilleur rappel à 50 % de précision (R@P50)
- Ces trois vecteurs de prédictions (Alpha, Beta et Gamma) ont ensuite été enregistrés pour être utilisés comme caractéristiques candidates dans le modèle multimodal présenté dans la partie 03.

Sélection de l'architecture du modèle TDM

Avant de procéder à l'analyse de fusion finale, une expérience préliminaire a été réalisée pour sélectionner l'architecture optimale du modèle TDM. Deux versions distinctes du modèle TDM (V1 et V2) ont été développées. Ces versions différaient par la conception de leur couche de projection (V2 utilisant une simple projection linéaire) et ont été entraînées avec différents ensembles d'hyperparamètres. Les deux versions ont été entraînées indépendamment en utilisant la méthodologie de validation croisée à 5 plis décrite précédemment. Les performances des six modèles TDM unimodaux résultants (c'est-à-dire V1-alpha, V1-beta, V1-gamma, V2-alpha, V2-beta, et V2-gamma) ont ensuite été comparées.

L'architecture V2 a été sélectionnée pour l'analyse multimodale finale. Étant donné que le point de contrôle V2-alpha a obtenu l'AUROC OOF la plus élevée parmi toutes les variantes de modèle, l'ensemble des points de contrôle V2 (alpha, beta et gamma) a été conservé comme caractéristiques d'imagerie pour tous les modèles en aval.

Partie 2 : Le modèle bioclinique

Architecture du modèle

Un classificateur XGBoost (Extreme Gradient Boosting) [13] a été choisi comme modèle bioclinique en raison de ses hautes performances, de sa rapidité et de sa gestion robuste des données tabulaires.

Un choix méthodologique délibéré a été fait de ne pas effectuer d'optimisation des hyperparamètres (HPO) pour les modèles

XGBoost. Étant donné la taille limitée de la cohorte (n=255) et le faible nombre d'événements positifs (n=42), toute tentative d'HPO aurait présenté un risque élevé de sur-ajustement et de découverte de résultats optimaux non reproductibles.

Pour éviter ce biais, nous avons privilégié une approche "plug-and-play" en utilisant un modèle XGBoost standardisé (avec `n_estimators=100`) pour toutes les comparaisons. Cette méthodologie nous permet de tester la **robustesse** et la **force intrinsèque du signal** de chaque modalité, répondant ainsi à notre question de recherche sur l'utilité pratique et "prête à l'emploi" de la fusion multimodale, plutôt que de poursuivre une performance maximale théorique qui ne serait probablement pas généralisable.

Pour assurer une évaluation robuste et maintenir la cohérence avec la génération de caractéristiques OOF du modèle TDM, les modèles biocliniques ont également été entraînés et évalués en utilisant le partitionnement de validation croisée à 5 plis des données d'entraînement.

Pour faire face au déséquilibre significatif des classes dans l'ensemble de données (16,1 % de PAS), le classificateur XGBoost a été configuré avec le paramètre `scale_pos_weight`. Ce paramètre a été calculé dynamiquement pour chaque pli d'entraînement comme le rapport des instances négatives (non-PAS) aux instances positives (PAS), garantissant une pénalité équilibrée pour la mauvaise classification.

Le modèle a été entraîné en utilisant l'ensemble complet des caractéristiques traitées et conçues, tirant parti de la capacité

native de XGBoost à effectuer une sélection interne des caractéristiques.

Prétraitement des données et ingénierie des caractéristiques

Pour assurer l'intégrité méthodologique et prévenir la fuite de données, toutes les étapes de prétraitement ont été encapsulées dans un Pipeline scikit-learn et un ColumnTransformer. Cela garantit que toutes les statistiques (par exemple, les médianes de l'imputer, les moyennes du scaler) ont été apprises uniquement sur les données d'entraînement de chaque pli de validation croisée.

- **Exclusion de caractéristiques :**
 - Les caractéristiques présentant un taux élevé de données manquantes (>25 %, par exemple, Bun, PDW, HDL, cpr, PAB) ont été exclues de l'analyse.
 - Les scores de gravité établis (MCTSI, BISAP, Ranson, SABP) ont été exclus de l'ensemble de caractéristiques. Cela nous a permis d'évaluer la quantité d'informations prédictives pouvant être apprises directement à partir des caractéristiques cliniques brutes sans dépendre de scores pré-calculés.
- **Stratégie Imputer-puis-Calculer :** Pour éviter le biais de l'imputer dû à des caractéristiques redondantes, toutes les caractéristiques pré-calculées dans l'ensemble de données brutes (par exemple, MAP, nlr, dtr, tyg, car, rcr) ont été supprimées. Leurs caractéristiques d'entrée brutes (par exemple, SBP, DBP, NEUT, TbiL) ont été conservées. Après imputation, ces six caractéristiques ont été recalculées et transmises au modèle en tant que caractéristiques conçues.
- **Pipeline de caractéristiques numériques :** Un pipeline en 02 étapes a été appliqué à

toutes les colonnes numériques :

1. Met à l'échelle les caractéristiques numériques avec *RobustScaler*.
2. Impute les valeurs manquantes avec *KNNImputer* (3 voisins).
3. Déséchelonne les données.
4. Ajoute des caractéristiques calculées à partir des données existantes.

- **Pipeline de caractéristiques catégorielles** : Un pipeline en 2 étapes a été appliqué à toutes les colonnes catégorielles (par exemple, Sexe, tabac, EP, SIRS) :

1. Impute les valeurs manquantes avec *SimpleImputer* (stratégie « most frequent »).
2. Encode les colonnes binaires avec *OneHotEncoder* (sans redondance)

Partie 3 : Le modèle de fusion multimodale

Stratégie de fusion

La dernière étape de la méthodologie a consisté à intégrer les caractéristiques d'imagerie (de la partie 01) avec les données biocliniques (de la partie 02). Les **scores de prédiction hors échantillon (OOF)** des trois points de contrôle **CT Model V2** sélectionnés (alpha, beta, et gamma) ont été concaténés en tant que nouvelles caractéristiques individuelles au jeu de données biocliniques.

Cela a créé trois jeux de données multimodaux distincts :

1. **Bioclinique + Score CT alpha**
2. **Bioclinique + Score CT beta**
3. **Bioclinique + Score CT gamma**
4. **Entraînement**

Les mêmes divisions de validation croisée à 5 plis et procédures d'entraînement de la partie 2 ont été appliquées. Le traitement

des caractéristiques biocliniques est identique à la partie 02 (bioclinique pur), et la nouvelle caractéristique de score CT a été transmise sans traitement au modèle pour la classification.

Cette méthodologie a abouti à un total de **quatre modèles** pour comparaison :

- **Référence** : Bioclinique uniquement
- **Multimodal alpha** : (Bioclinique + Score V2-alpha)
- **Multimodal beta** : (Bioclinique + Score V2- beta)
- **Multimodal gamma** : (Bioclinique + Score V2- gamma)

Les quatre modèles ont été entraînés et évalués en utilisant les mêmes divisions de validation croisée à 5 plis pour assurer une comparaison équitable.

Analyse statistique et évaluation

Toutes les évaluations des modèles ont été effectuées dans le cadre de la validation croisée à 5 plis, aucun ensemble de test indépendant n'a été utilisé.

Les métriques de performance (AUROC, AUPRC, Rappel à 50 % de précision) ont été calculées de deux manières :

1. **Par pli** : La métrique a été calculée sur l'ensemble de validation de chaque pli. La performance finale rapportée pour la robustesse du modèle est la moyenne et l'écart type de ces scores à 5 plis.
2. **OOF agrégé** : Les prédictions des 5 plis de validation ont été concaténées pour former un seul vecteur de prédiction "hors échantillon" (OOF) pour l'ensemble du jeu de données. Les métriques de performance globales ont ensuite été calculées une fois sur ces scores agrégés.

Pour déterminer si l'ajout de données TDM apportait une amélioration statistiquement significative, l'AUROC OOF agrégée de chaque modèle multimodal (alpha, beta, gamma) a été comparée à l'AUROC OOF agrégée de la référence "Bioclinique uniquement". Cette comparaison a été effectuée à l'aide du test de DeLong, avec des résultats rapportés à l'aide d'intervalles de confiance (IC) à 95 % et de valeurs p.

Résultats

L'évaluation de notre pipeline multimodal

s'est déroulée en trois étapes successives : en premier l'analyse des modèles d'imagerie CT, puis l'évaluation du modèle bioclinique de référence, et enfin la fusion multimodale. Nous présentons les résultats dans cet ordre logique, en distinguant systématiquement les performances moyennes par pli (reflétant la variabilité locale) et les performances OOF agrégées (reflétant la généralisation globale).

Deux architectures CT (V1 et V2) avec trois critères d'optimisation (Alpha, Beta, Gamma) ont été comparées.

Tableau 01 : Performances OOF des modèles TDM V1 et V2.

Modèle	Critère de sélection	AUC OOF	AUPRC OOF	Rappel OOF@P50
Alpha-V1	Meilleur par AUC	0.7569	0.3632	0.5238
Beta-V1	Meilleur par AUPRC	0.7514	0.4086	0.3095
Gamma-V1	Meilleur par Rappel@P50	0.7440	0.4693	0.4524
Alpha-V2	Meilleur par AUC	0.7748	0.4177	0.4048
Beta-V2	Meilleur par AUPRC	0.7281	0.3409	0.2857
Gamma-V2	Meilleur par Rappel@P50	0.7216	0.3770	0.3810

Synthèse Tableau 01: V2-Alpha atteint l'AUROC OOF la plus élevée (0,7748), justifiant la sélection de l'ensemble V2 pour la fusion multimodale.

Fort de cette sélection architecturale, les performances des trois variantes V2 ont été analysées pour comprendre leur comportement en termes de généralisation.

Tableau 02 : Résultats par pli : modèle TDM V2-Alpha (meilleur AUC par pli).

Pli	AUC	AUPRC	Rappel@P50
1	0.7936	0.4160	0.6250
2	0.8721	0.4520	0.8750
3	0.9215	0.7079	0.8750
4	0.7143	0.2735	0.0000
5	0.7937	0.4206	0.1111
Moyenne	0.8190	0.4540	0.4972
Écart-type	0.0715	0.1412	0.3736

Synthèse Tableau 02 : Alpha démontre une performance moyenne par pli de 0,8190 en AUC, 0,4540 en AUPRC et 0,4972 en

Recall@P50, avec une performance OOF agrégée conservant 94,6% de l'AUC moyenne (0,7748 OOF).

Tableau 03 : Résultats par pli : modèle TDM V2-Beta (meilleur AUPRC par pli).

Pli	AUC OOF	AUPRC OOF	Rappel OOF@P50
1	0.7500	0.5262	0.7500
2	0.8343	0.5455	0.5000
3	0.8401	0.7158	0.7500
4	0.7063	0.2823	0.0000
5	0.7937	0.6592	0.6667
Moyenne	0.7849	0.5458	0.5333
Écart-type	0.0509	0.1494	0.2819

Synthèse Tableau 03 : Beta atteint une performance moyenne par pli de 0,7849 en AUC, 0,5458 en AUPRC (la meilleure) et 0,5333 en Recall@P50, mais montre une

dégradation sévère en OOF avec seulement 62,5% de conservation pour l'AUPRC (0,3409 OOF).

Tableau 04 : Résultats par pli : modèle TDM V2-Gamma (meilleur R@P50 par pli).

Pli	AUC OOF	AUPRC OOF	Rappel OOF@P50
1	0.7500	0.5262	0.7500
2	0.8488	0.5208	0.8750
3	0.9215	0.7079	0.8750
4	0.6614	0.2547	0.1111
5	0.7937	0.6592	0.6667
Moyenne	0.7951	0.5337	0.6556
Écart-type	0.0881	0.1576	0.2835

Synthèse Tableau 04: Gamma obtient une performance moyenne par pli de 0,7951 en AUC, 0,5337 en AUPRC et 0,6556 en Recall@P50 (la meilleure), avec une conservation OOF modérée de 58,1% pour le Recall@P50 (0,3810 OOF). Ces résultats

(tableaux 2 - 4) révèlent qu'Alpha maintient la meilleure généralisation OOF malgré des performances par pli parfois inférieures à Beta et Gamma sur leurs métriques respectives. Nous comparons maintenant ces modèles CT au modèle bioclinique de référence.

Tableau 05 : Modèle bioclinique : performance par pli et OOF global.

Plis (Fold)	AUC	AUPRC	R@P50
1	0.816860	0.494672	0.375000
2	0.720930	0.421408	0.250000
3	0.808140	0.631588	0.750000

4	0.920635	0.794302	0.888889
5	0.820106	0.521223	0.555556
Moyenne	0.8173	0.5726	0.5639
Écart-type	0.0708	0.1451	0.2620
Global OOF	0.8196	0.5566	0.4048

Synthèse Tableau 05 : Le modèle bioclinique atteint une performance moyenne par pli de 0,8173 en AUC, 0,5726 en AUPRC et 0,5639 en Recall@P50, avec une excellente conservation OOF (100,3% pour l'AUC : 0,8196 OOF ; 97,2% pour l'AUPRC : 0,5566 OOF), démontrant un

apprentissage robuste sans surapprentissage significatif.

Cette baseline performante et stable (0,82 AUC) constitue un standard exigeant. Nous évaluons maintenant si l'ajout des caractéristiques CT peut apporter une amélioration significative par fusion multimodale.

Tableau 06 : Modèle Multimodal Alpha (TDM V2-Alpha + Bioclinique).

Plis (Fold)	AUC	AUPRC	R@P50
1	0.784884	0.524276	0.375000
2	0.767442	0.572320	0.500000
3	0.863372	0.669777	0.875000
4	0.941799	0.882496	0.888889
5	0.849206	0.609452	0.777778
Moyenne	0.8413	0.6517	0.6833
Écart-type	0.0694	0.1396	0.2327
Global OOF	0.8492	0.6398	0.6190

Synthèse Tableau 06 : MM-Alpha atteint une performance moyenne par pli de 0,8413 en AUC, 0,6517 en AUPRC et 0,6833 en Recall@P50, avec une performance OOF

globale de 0,8492 en AUC (conservation 100,9%), 0,6398 en AUPRC et 0,6190 en Recall@P50, une amélioration substantielle par rapport au baseline bioclinique.

Tableau 07 : Modèle Multimodal Beta (TDM V2-Beta + Bioclinique).

Plis (Fold)	AUC	AUPRC	R@P50
1	0.845930	0.535340	0.250000
2	0.723837	0.407452	0.250000
3	0.805233	0.649475	0.750000
4	0.888889	0.767910	0.888889
5	0.843915	0.665095	0.555556
Moyenne	0.8216	0.6051	0.5389
Écart-type	0.0621	0.1378	0.2891
Global OOF	0.8186	0.5656	0.3810

Synthèse Tableau 07 : MM-Beta atteint une performance moyenne par pli de 0,8216 en AUC, 0,6051 en AUPRC et 0,5389 en Recall@P50, avec une performance OOF

globale de 0,8186 en AUC, essentiellement identique au baseline bioclinique (0,8196).

Tableau 08 : Modèle Multimodal Gamma (TDM V2-Gamma + Bioclinique).

Plis (Fold)	AUC	AUPRC	R@P50
1	0.796512	0.521501	0.250000
2	0.857558	0.552701	0.375000
3	0.831395	0.648630	0.750000
4	0.888889	0.774502	0.777778
5	0.862434	0.618852	0.333333
Moyenne	0.8474	0.6232	0.4972
Écart-type	0.0350	0.0986	0.2478
Global OOF	0.8391	0.5658	0.4762

Synthèse Tableau 08 : MM-Gamma atteint une performance moyenne par pli de 0,8474 en AUC, 0,6232 en AUPRC et 0,4972 en Recall@P50, avec une performance OOF globale de 0,8391 en AUC une amélioration non statistiquement significative par rapport

au baseline. Pour déterminer la significativité statistique de ces améliorations apparentes, nous avons appliqué le test de DeLong comparant les AUROC OOF de chaque modèle multimodal au baseline bioclinique.

Tableau 09 : Test de DeLong : modèles multimodaux vs. modèle bioclinique.

Modèle	AUC Unimodal	AUC MM	ΔAUC	z-score	p-value	IC 95% MM	IC 95% Bioc
MM Alpha	0.8196	0.8492	0.0296	-2.0184	<u>0.0435</u>	(0.7866, 0.9118)	(0.7503, 0.8888)
MM Beta	0.8196	0.8186	-0.0010	0.0696	0.9445	(0.7504, 0.8868)	(0.7503, 0.8888)
MM Gamma	0.8196	0.8391	0.0196	-0.9744	0.3299	(0.7771, 0.9012)	(0.7503, 0.8888)

Synthèse Tableau 09 : Seul MM-Alpha démontre une amélioration statistiquement significative (ΔAUC +0,0296, p=0,0435, IC 95% [0,7866-0,9118]), tandis que MM-Beta (ΔAUC -0,0010, p=0,9445) et MM-Gamma (ΔAUC +0,0196, p=0,3299) n'atteignent pas la significativité statistique.

Ces résultats établissent que la fusion multimodale avec le score CT V2-Alpha améliore significativement la prédiction de sévérité de la pancréatite aiguë, avec un impact clinique majeur sur la détection des cas sévères (Recall@P50 : +52,9%).

Figure 01 : Courbes ROC comparant les performances du modèle bioclinique et des trois modèles multimodaux.

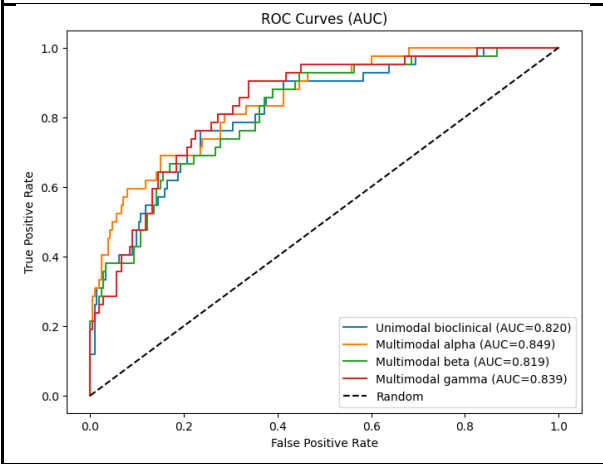
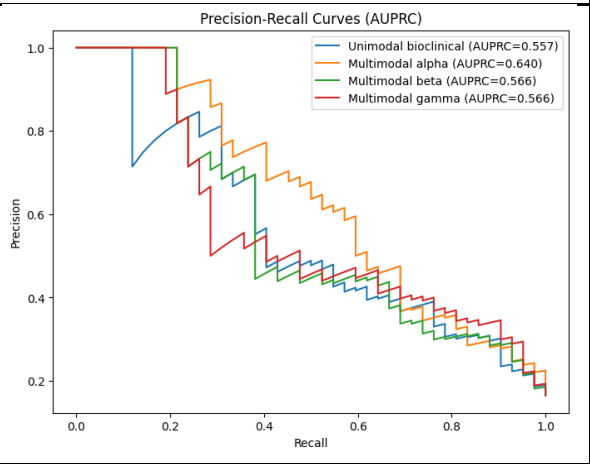


Figure 02 : Courbes AUPRC comparant les performances du modèle bioclinique et des trois modèles multimodaux.



Les figures 01 et 02 : viennent compléter visuellement les résultats présentés dans les tableaux 05 à 08, en illustrant

respectivement les courbes ROC et AUPRC obtenues en validation out-of-fold (OOF).

Discussion

Les pancréatites aiguës sévères représentent environ 10 à 20% des cas avec une mortalité pouvant atteindre 30%, nécessitant une identification précoce des patients à risque. Les scores traditionnels (APACHE II, BISAP) présentent plusieurs limitations : aucun système de score n'est idéal avec une sensibilité et spécificité élevées.

L'objectif principal de notre étude est de pouvoir déterminer un modèle d'IA multimodal, combinant des données biocliniques et des caractéristiques d'imagerie TDM, pouvait améliorer la prédiction de la gravité de la pancréatite aiguë par rapport à un modèle unimodal bioclinique. Les résultats obtenus confirment cette hypothèse : l'intégration de caractéristiques TDM, extraites par un réseau ResNet-50 et optimisées par un entraînement à double tâche, améliore la performance prédictive globale.

Analyse des modèles CT

Le modèle V2-Alpha obtient la meilleure performance OOF (0,7748 AUC) parmi six variantes testées, justifiant la sélection de l'ensemble V2 pour la fusion multimodale. Chaque modèle excelle pour sa métrique cible en moyenne par pli : Alpha pour AUC (0,8190), Beta pour AUPRC (0,5458), Gamma pour Recall@P50 (0,6556). Les variations importantes (écart-type >0,25) reflètent la difficulté d'entraîner sur petits groupes déséquilibrés (~8-9 cas sévères/groupe).

L'évaluation agrégée out-of-fold (OOF), qui regroupe toutes les prédictions (scores entre 0 et 1) faites par le modèle sur des patients jamais vus pendant son entraînement, constitue un test essentiel de généralisation. Contrairement à la validation par groupe (par pli) qui évalue la performance sur de petits sous-ensembles, l'OOF teste si le

modèle maintient sa capacité de distinction lorsqu'on agrège les prédictions sur l'ensemble des 255 patients. Un modèle qui conserve une bonne performance en OOF a appris des caractéristiques réellement prédictives de la sévérité, plutôt que des particularités propres à chaque groupe. Le pourcentage de performance conservé (OOF/moyenne par groupe) indique le degré de généralisation.

L'agrégation OOF révèle un renversement majeur : Alpha domine les trois métriques avec une excellente conservation de performance (AUC: 94,6%, AUPRC: 92,0%, Recall@P50: 81,4%). À l'inverse, Beta et Gamma montrent une dégradation sévère (conservation 53-71% pour leurs métriques cibles), suggérant un surapprentissage aux particularités de chaque groupe d'entraînement. La hiérarchie finale obtenue se résume : Alpha > Gamma > Beta. La performance finale d'Alpha : AUC 0,7748 , AUPRC 0,4177.

Analyse du modèle Bioclinique

Le modèle bioclinique XGBoost non optimisé (n_estimators=100) atteint des performances OOF remarquables : AUC 0,8196, AUPRC 0,5566, Recall@P50 0,4048. Sa généralisation est exceptionnelle avec une conservation de performance >97% pour AUC et AUPRC entre évaluation par groupe et OOF (100,3% et 97,2% respectivement), contrastant fortement avec la dégradation observée pour les modèles CT Beta et Gamma (conservation 53-71%). Cette stabilité démontre l'apprentissage de patterns biocliniques robustes sans surapprentissage significatif. La performance est d'autant plus remarquable que les scores de gravité validés (Ranson, BISAP, MCTSI) ont été délibérément exclus : le modèle atteint 0,82 AUC uniquement avec des données

biocliniques brutes collectées à l'admission (24h), exploitant efficacement sa capacité de sélection interne de caractéristiques. La variabilité inter-groupes observée (AUC variant de 0,72 à 0,92, avec le Pli 4 exceptionnellement performant à 0,92), est cohérente avec la taille limitée de la cohorte (255 patients, 42 cas sévères). Ce modèle "prêt à l'emploi" établit ainsi une baseline robuste et généralisable qui pourrait rivaliser avec les systèmes de scoring traditionnels tout en offrant une prédiction précoce dès 24h, constituant un standard de performance élevé et exigeant pour notre évaluation multimodale.

Analyse de la fusion multimodale

Le résultat principal réside dans l'amélioration par fusion multimodale (Tableaux 5-9), interprétée à la lumière de deux observations clés :

- CT-Alpha a démontré une supériorité OOF dans les trois métriques,
- le baseline bioclinique établit déjà un standard élevé (0,8196 AUC) avec excellente rétention (97-100%).

MM-Alpha atteint une AUROC OOF de 0,8492 vs 0,8196 (baseline), soit Δ AUC +0,0296 ($p=0,0435$, test de DeLong), avec IC 95% [0,7866-0,9118] décalé positivement vs baseline [0,7503-0,8888]. Cette amélioration, bien que numériquement modeste, est remarquable car elle surpasse une baseline déjà performante sans optimisation des hyperparamètres XGBoost, démontrant que le signal V2-Alpha apporte une information complémentaire immédiatement exploitable. L'amélioration s'étend à toutes les métriques : AUPRC 0,6398 vs 0,5566 (Δ +0,0832, +14,9%) et surtout Recall@P50 0,6190 vs 0,4048 (Δ +0,2142, +52,9%).

Pour l'impact clinique, à précision fixe de 50%, MM-Alpha détecte 62% des cas

sévères vs 40% pour le baseline, correspondant à 09 patients PAS supplémentaires correctement identifiés dans notre cohorte (26/42 vs 17/42), une amélioration transformatrice pour l'identification précoce et l'admission en soins intensifs. La fusion améliore également la stabilité : rétention per-fold/OOF passe de 71,8% à 90,6% pour Recall@P50 (+18,8 points), indiquant que le signal CT stabilise la détection des cas sévères.

Le MM-Beta (AUC 0,8186, Δ AUC -0,0010, $p=0,9445$) et le MM-Gamma (AUC 0,8391, Δ AUC +0,0196, $p=0,3299$) n'améliorent pas significativement le baseline. Cette non-significativité reflète leurs faibles performances CT unimodales : V2-Beta (AUC 0,7281, rétention AUPRC 62,5%) et V2-Gamma (AUC 0,7216, rétention R@P50 58,1%) ont montré une dégradation sévère suggérant surapprentissage. Gamma montre néanmoins une tendance positive ($p=0,33$), reflétant sa hiérarchie OOF supérieure à Beta. Ces signaux, optimisés pour des métriques sensibles aux seuils (AUPRC, Recall@P50), peuvent représenter des informations "complexes" ou "spécialisées" nécessitant une calibration des poids ou une optimisation des hyperparamètres que le XGBoost standardisé n'a pas su exploiter. L'hypothèse reste que dans un cadre pleinement optimisé sur cohortes plus larges, ces signaux, particulièrement Gamma, pourraient apporter une valeur additionnelle.

Le succès de MM-Alpha valide l'hypothèse centrale : les informations morphologiques (TDM) et systémiques (biocliniques) sont complémentaires. Les données biocliniques à 24h reflètent la réponse systémique précoce (inflammation, dysfonction

organique), tandis que la TDM capture les changements morphologiques (nécrose, collections, thrombose vasculaire) pouvant précéder la détérioration biologique. Le classificateur XGBoost a spontanément sélectionné le "Score V2-Alpha" comme caractéristique précieuse parmi les variables biocliniques, démontrant son apport non-redondant. L'amélioration massive du Recall@P50 (+53%) suggère que la TDM aide particulièrement à identifier les cas sévères manqués par les seuls critères biocliniques.

Limites de l'étude

- **Taille de la cohorte** : la cohorte monocentrique limitée (255 patients, 42 cas sévères) avec risque de surapprentissage substantiel [8]. Chaque pli d'entraînement contient ~34 cas sévères et chaque pli de validation ~8 cas sévères, insuffisant pour optimisation stable de métriques complexes. Cette limitation explique la dégradation sévère observée pour Beta/Gamma (conservation 53-71%) et limite la généralisabilité des résultats.
- **Absence de validation externe** : l'évaluation réalisée uniquement via validation croisée 5-plis, sans ensemble de test externe indépendant. La performance réelle des modèles sur une cohorte prospective ou provenant d'une autre institution reste inconnue. Une validation multicentrique est nécessaire pour confirmer ces résultats prometteurs.
- **Approche "plug-and-play" non optimisée** : le choix méthodologique délibéré de ne pas optimiser les hyperparamètres XGBoost pour éviter le surapprentissage sur petite cohorte. Cette approche présente un double

enjeu : d'une part, l'optimisation des hyperparamètres pourrait améliorer la performance du modèle bioclinique seul, créant ainsi un baseline encore plus exigeant qui pourrait masquer l'amélioration apportée par la fusion multimodale ; d'autre part, sur notre cohorte limitée (255 patients, 42 cas sévères), une telle optimisation risquerait un surapprentissage significatif, produisant des résultats non reproductibles. Nous avons donc privilégié un modèle par défaut non optimisé, potentiellement moins performant, pour tester l'apport immédiat brut de l'information TDM dans un contexte "prêt à l'emploi". Cette stratégie évalue la valeur ajoutée minimale garantie de la multimodalité. Des travaux futurs sur cohortes plus larges (n>500) permettront de construire des modèles CT plus performants avec davantage d'exemples d'entraînement, et d'explorer l'étendue de la performance maximale via optimisation complète des hyperparamètres. XGBoost révélerait ainsi le potentiel complet de la fusion multimodale, y compris pour les variantes Beta et Gamma actuellement non-significatives.

- **Interprétabilité limitée** : Le score TDM reste une "boîte noire". Bien que la fusion améliore les métriques, les régions anatomiques ou caractéristiques morphologiques spécifiques responsables de la prédiction restent inconnues. Les travaux futurs pourraient utiliser des méthodes d'explicabilité (Grad-CAM, SHAP) pour visualiser les zones d'intérêt et améliorer la confiance clinique.

Des études récentes ont démontré la valeur des approches d'IA multimodales pour la prédiction de la sévérité de la PA. Xiang et Shang [14] ont développé un modèle d'apprentissage profond multi-vues combinant indices cliniques et images TDM, atteignant une précision de 80,26% (IC 95% : 79,58%–80,94%) dans une cohorte de 772 patients. Yin et al [1] ont créé PrismSAP, un modèle multimodal fusionnant dossiers médicaux électroniques, radiomique 3D-TDM et caractéristiques d'apprentissage profond 2D-TDM sur 1 221 patients de plusieurs centres, atteignant une AUC de validation externe de 0,916 (0,873–0,960) pour la prédiction de la PAS, surpassant tous les modèles unimodaux et les systèmes de scores traditionnels. De même, Qi et Lu [15] ont combiné radiomique TDM et paramètres cliniques (calcium sérique, CRP, leucocytes) chez 245 patients, atteignant des AUC de 0,905 (entraînement) et 0,908 (validation) pour la prédiction MSAP/PAS.

Conclusion

Notre étude démontre avec succès qu'un modèle d'IA multimodal fusionnant données biocliniques et caractéristiques TDM extraites par apprentissage profond surpasse significativement un modèle bioclinique seul pour la prédiction de la sévérité de la pancréatite aiguë. Le modèle MM-Alpha atteint une amélioration statistiquement significative de l'AUROC (0,8492 vs 0,8196, $p=0,0435$) et surtout une augmentation remarquable de 53% du Recall@P50, permettant l'identification correcte de 9 patients PAS supplémentaires dans notre cohorte, une avancée cliniquement transformatrice pour la

détection précoce et l'orientation vers les soins intensifs.

Le succès de MM-Alpha valide la complémentarité des informations morphologiques (TDM) et systémiques (biocliniques) : la TDM capture des changements structuraux précoces qui précèdent ou complètent les marqueurs biologiques, permettant d'identifier des cas sévères manqués par les seuls critères biocliniques.

Ces résultats prometteurs ouvrent la voie à des outils d'aide à la décision clinique pour la stratification précoce des risques en pancréatite aiguë. Cependant, plusieurs étapes essentielles restent nécessaires avant l'implémentation clinique :

1. Validation externe sur cohortes multicentriques et prospectives ($n>500$) pour confirmer la généralisabilité,
2. Exploration du potentiel maximal via optimisation des hyperparamètres et architectures avancées,
3. Amélioration de l'interprétabilité par méthodes d'explicabilité (Grad-CAM) pour identifier les biomarqueurs morphologiques prédictifs.
4. Evaluation de l'impact clinique réel sur les décisions thérapeutiques, la durée de séjour et la mortalité.

L'approche multimodale proposée établit néanmoins un cadre méthodologique robuste et reproductible, adaptable à d'autres pathologies où la fusion d'informations cliniques et d'imagerie pourrait améliorer la prédiction pronostique.

Remerciements

Nous tenons à remercier chaleureusement les médecins du Service universitaire de chirurgie générale de l'EPH d'Aïn Taya pour la valeur de leur expertise médicale.

Nous exprimons également notre profonde gratitude à l'équipe de l'École Nationale Supérieure d'Informatique ESI pour sa contribution dans le domaine de l'intelligence artificielle. Enfin, nous remercions l'ensemble de l'équipe de PrismSAP [1] pour la mise à disposition des données nécessaires à la réalisation de cette étude.

Références bibliographiques

- [1]- Yin M, Lin J, Wang Y, Liu Y, Zhang R, Duan W, et al. Development and validation of a multimodal model in predicting severe acute pancreatitis based on radiomics and deep learning. *Int J Med Inf.* 1 avr 2024;184:105341.
- [2]- Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus. *Gut.* 2013;62(1):102-11.
- [3]- JH R. Prognostic signs and the role of operative management in acute pancreatitis. *Surg Gynecol Obstet.* 1974;139:69-81.
- [4]- Singh VK, Wu BU, Bollen TL, Repas K, Maurer R, Johannes RS, et al. A prospective evaluation of the bedside index for severity in acute pancreatitis score in assessing mortality and intermediate markers of severity in acute pancreatitis. *Off J Am Coll Gastroenterol ACG.* 2009;104(4):966-71.
- [5]- APACHE I. Apache ii: a severity of disease classification system. 1985;
- [6]- Morteale KJ, Wiesner W, Intriere L, Shankar S, Zou KH, Kalantari BN, et al. A modified CT severity index for evaluating acute pancreatitis: improved correlation with patient outcome. *Am J Roentgenol.* 2004;183(5):1261-5.
- [7]- Qian R, Zhuang J, Xie J, Cheng H, Ou H, Lu X, et al. Predictive value of machine learning for the severity of acute pancreatitis: A systematic review and meta-analysis. *Heliyon.* 2024;10(8).
- [8]- López Gordo S, Ramirez-Maldonado E, Fernandez-Planas MT, Bombuy E, Memba R, Jorba R. AI and Machine Learning for Precision Medicine in Acute Pancreatitis: A Narrative Review. *Medicina (Mex).* 2025;61(4):629.
- [9]- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics.* sept 1988;44(3):837.
- [10]- He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Las Vegas, NV, USA: IEEE; 2016 [cité 2 sept 2025]. p. 770-8. Disponible sur: <http://ieeexplore.ieee.org/document/7780459/>
- [11]- Zheng Q, Zhao W, Wu C, Zhang X, Dai L, Guan H, et al. Large-scale long-tailed disease diagnosis on radiology images. *Nat Commun.* 2024;15(1):10147.
- [12]- Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised

contrastive learning. Adv Neural Inf Process Syst. 2020;33:18661-73.

[13]- Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-94.

[14]- Xiang K, Shang D. Construction of a Multi-View Deep Learning Model for the Severity Classification of Acute Pancreatitis. Discov Med. 2025;37(192):73.

[15]- Qi M, Lu C, Dai R, Zhang J, Hu H, Shan X. Prediction of acute pancreatitis severity based on early CT radiomics. BMC Med Imaging. 2024;24(1):321.