

**République Algérienne Démocratique et Populaire Ministère de
l'Enseignement Supérieur et de la Recherche Scientifique Université Amar
THELIDJI-Laghouat Faculté de Médecine Département de Médecin**



Validation d'un système d'aide à la décision de l'indication chirurgicale basé sur l'IA pour le triage diagnostique de la douleur abdominale aiguë

Mémoire de fin d'étude pour l'obtention du diplôme de docteur en médecine

Présentée par :

- **MOUAFKI Nouredine**

Devant le jury :

Dr.BOUTAGHENE Lila : Présidente

Pr.OUALID Mouhamed : Encadreur

Dr .BOUDIA Mokhtaria : Examineur

Année Universitaire : 2025/2026

DÉDICACES

Je dédie ce travail à:

Mes chers parents,

Pour leur amour inconditionnel, leurs sacrifices et leur soutien constant tout au long de mon parcours.

Ma famille,

Pour leur encouragement et leur présence à mes côtés.

Mes enseignants,

Qui m'ont transmis le savoir et la passion de la médecine.

Mes amis et collègues,

Pour les moments de partage et de solidarité.

Tous les patients,

Qui sont la raison d'être de notre vocation médicale

REMERCIEMENTS

Nous remercions tout d'abord **Allah le Tout-Puissant** de nous avoir donné la force, le courage et la patience pour mener à bien ce projet.

Nous tenons à exprimer notre profonde gratitude à toutes les personnes qui ont contribué à la réalisation de ce travail.

Nos sincères remerciements vont à notre directeur de thèse, **Professeur OUALID Mouhamed**, pour son encadrement au long de cette recherche.

Nous remercions les membres du jury d'avoir accepté d'évaluer ce travail et de nous faire bénéficier de leurs remarques constructives.

Enfin, nous remercions tous ceux qui, de près ou de loin, ont contribué à l'aboutissement de ce travail.

RÉSUMÉ

Introduction La douleur abdominale aiguë (DAA) représente 5 à 10 % des consultations dans les services d'urgence et constitue un défi diagnostique majeur en raison de la diversité des étiologies possibles, allant de pathologies bénignes auto-résolutives à des urgences chirurgicales mettant en jeu le pronostic vital [1,3]. Le retard au diagnostic d'une urgence chirurgicale est associé à une augmentation significative de la morbidité et de la mortalité, en particulier chez les sujets âgés ou immunodéprimés [7,15]. L'intelligence artificielle (IA) a connu un essor considérable en médecine, notamment à travers les systèmes d'aide à la décision clinique (*Clinical Decision Support Systems*, CDSS) [35]. Les modèles d'apprentissage automatique, entraînés sur des données cliniques, biologiques et d'imagerie, ont montré des performances prometteuses pour le diagnostic étiologique et la prédiction du recours à la chirurgie dans les douleurs abdominales aiguës, avec des aires sous la courbe comprises entre 0,85 et 0,95 [32,39]. Cependant, la majorité de ces modèles n'ont pas été validés en conditions réelles, leur explicabilité reste limitée et leur intégration clinique se heurte à des obstacles techniques et réglementaires [27,42,43].

Objectif: Développer et valider une application basée sur l'intelligence artificielle pour la décision de l'indication chirurgicale devant DAA.

Méthodes: Un total de 205 dossiers médicaux relatifs à des douleurs abdominales aiguës ont été recueillis au sein de l'Hôpital Mixte de Laghouat et du Centre Hospitalier Universitaire Mustapha Bacha, durant la période allant du 1er novembre 2025 au 12 janvier 2026, conformément au modèle opératoire « Annexe I ». Un tri de ces cas a ensuite été réalisé et tous les dossiers ne répondant pas aux critères d'inclusion prédéfinis pour l'étude ont été exclus. Au final, 87 dossiers médicaux ont été jugés éligibles pour l'analyse. Les données issues de ces dossiers ont été soumises au modèle Kashef AI, préalablement conçu et programmé. La valeur prédictive positive (VPP), La valeur prédictive négative (VPN), La sensibilité et la spécificité dudit modèle pour le diagnostic du degré d'urgence de la douleur abdominale ont alors été évaluées..

Résultats : 87 patients ont été recrutés et testés sur le modèle Kashif Ai. On a trouvé une sensibilité acceptable (75 %) et une excellente valeur prédictive négative (90 %), ce qui en fait

potentiellement un outil de tri primaire utile pour identifier les patients à faible risque d'urgence chirurgicale. En revanche, sa spécificité médiocre (50,7 %) et sa valeur prédictive positive effondrée (25,5 %) contre-indiquent formellement son utilisation comme outil de confirmation diagnostique.

Conclusion: L'application d'IA développée constitue un outil prometteur d'aide à la décision de l'indication chirurgicale devant DAA, pouvant améliorer la rapidité et la précision diagnostique aux urgences .

Mots-clés: Douleur abdominale aiguë, Intelligence artificielle, Diagnostic, Système d'aide à la décision clinique.

ABSTRACT

Introduction: Acute abdominal pain (AAP) accounts for 5–10% of emergency department visits and represents a major diagnostic challenge due to the wide range of possible aetiologies, from self-limiting benign conditions to life-threatening surgical emergencies [1,3]. Delayed diagnosis of a surgical emergency is associated with significantly increased morbidity and mortality, particularly in elderly or immunocompromised patients [7,15]. Artificial intelligence (AI) has expanded considerably in medicine, notably through Clinical Decision Support Systems (CDSS) [35]. Machine learning models trained on clinical, laboratory, and imaging data have shown promising performance for aetiological diagnosis and prediction of the need for surgery in acute abdominal pain, with areas under the curve ranging from 0.85 to 0.95 [32,39]. However, most of these models have not been validated in real-world settings, their explainability remains limited, and their clinical integration faces technical and regulatory barriers [27,42,43].

Objective: To develop and validate an artificial intelligence-based application for determining the need for surgical intervention in acute abdominal pain.

Methods: A total of 205 medical records of acute abdominal pain cases were collected at the Hôpital Mixte de Laghouat and the Centre Hospitalier Universitaire Mustapha Bacha, from 1 November 2025 to 12 January 2026, in accordance with the operative model “Appendix I”. These cases were then sorted and all records that did not meet the predefined inclusion criteria for the study were excluded. Ultimately, 87 medical records were deemed eligible for analysis. The data from these records were submitted to the previously designed and programmed Kashif AI model. The sensitivity, specificity, PPV and NPV of the model for diagnosing the degree of urgency of abdominal pain were then evaluated.

Results : Eighty-seven patients were recruited and tested with the Kashif AI model. An acceptable sensitivity (75%) and an excellent negative predictive value (90%) were found, potentially making it a useful primary triage tool to identify patients at low risk of surgical emergency. In contrast, its poor specificity (50.7%) and very low positive predictive value (25.5%) formally contraindicate its use as a diagnostic confirmation tool.

Conclusion: The developed AI application constitutes a promising decision-support tool for surgical indication in acute abdominal pain, capable of improving diagnostic speed and accuracy in the emergency department.

Keywords: Acute abdominal pain, Artificial intelligence, Diagnosis, Clinical decision support system.

TABLE DES MATIÈRES

DÉDICACES	i
REMERCIEMENTS	ii
RÉSUMÉ	iii
ABSTRACT	vi
TABLE DES MATIÈRES	vii
LISTE DES TABLEAUX	viii
LISTE DES FIGURES	ix
LISTE DES ABRÉVIATIONS	x
INTRODUCTION GÉNÉRALE	1
PARTIE I: CADRE THÉORIQUE CHAPITRE 1: LA DOULEUR ABDOMINALE AIGUË	3
1.1 Définition et concepts fondamentaux	3
1.2 Classification de la douleur abdominale	4
1.3 Épidémiologie.....	5
1.4 Étiologies principales.....	6
1.5 Approche diagnostique conventionnelle	7
CHAPITRE 2: PARAMÈTRES CLINIQUES ET BIOLOGIQUES	10
2.1 Examen clinique.....	10
2.2 Paramètres biologiques	11
2.3 Scores diagnostiques existants.....	13
CHAPITRE 3: INTELLIGENCE ARTIFICIELLE EN MÉDECINE	16
3.1 Fondements de l'IA	16

3.2 Applications médicales de l'IA	17
3.3 Systèmes d'aide à la décision clinique (CDSS)	18
3.4 Architecture et conception des grands modèles de langage (LLMs) pour le diagnostic médical	20
CHAPITRE 4: ÉTAT DE L'ART	24
4.1 Revue systématique de la littérature	24
4.2 Modèles IA existants.....	25
4.3 Limites et perspectives.....	26
PARTIE II: ÉTUDE PRATIQUE CHAPITRE 5: MATÉRIEL ET MÉTHODES	29
5.1 Type et cadre de l'étude	29
5.2 Population et échantillonnage.....	30
5.3 Variables de l'étude.....	31
5.4 Développement du modèle IA	33
CHAPITRE 6: RÉSULTATS	50
6.1 Caractéristiques de la population.....	50
6.2 Performance des modèles.....	52
CHAPITRE 7: DISCUSSION	56
7.1 Synthèse des résultats	56
7.2 Comparaison avec la littérature	57
7.3 Forces et limites.....	59
7.4 Implications Clinique.....	61
7.5 Perspectives.....	63
CONCLUSION GÉNÉRALE	73
RÉFÉRENCES BIBLIOGRAPHIQUES	74
ANNEXES	76

LISTE DES TABLEAUX

Tableau 1.1: Classification des douleurs abdominales selon la localisation.....	4
Tableau 2.1: Signes cliniques spécifiques et pathologies évoquées	11
Tableau 2.3: Score d'Alvarado pour l'appendicite	13
Tableau 4.1: Études récentes sur l'IA et la douleur abdominale	25
Tableau 6.1 : L'analyse de la répartition des niveaux de triage.....	50
Tableau 6.2 : La matrice de confusion 2×2, comparant le test de triage binaire (N1/N2 = Positif, N3 = Négatif) au diagnostic de référence.....	52
Tableau 6.3 : l'ensemble des indicateurs de performance.....	55

LISTE DES FIGURES

Figure 6.1 : Interface et module de localisation de la douleur.....	.67
Figure 6.2 : Module d'informations cliniques.....	.68
Figure 6.3 : Module antécédents et données biologiques.....	.69
Figure 6.4 : Module de résultats du diagnostic.	71

LISTE DES ABRÉVIATIONS

AAP : Acute Abdominal Pain (Douleur Abdominale Aiguë)

AUC : Area Under the Curve (Aire Sous la Courbe)

CDSS : Clinical Decision Support System

CNN : Convolutional Neural Network

CRP : C-Reactive Protein (Protéine C-Réactive)

DAA : Douleur Abdominale Aiguë

DL : Deep Learning (Apprentissage Profond)

FID : Fosse Iliaque Droite

FIG : Fosse Iliaque Gauche

GB : Globules Blancs (Leucocytes)

IA : Intelligence Artificielle

ML : Machine Learning (Apprentissage Automatique)

NFS : Numération Formule Sanguine

PCT : Procalcitonine

PNN : Polynucléaires Neutrophiles

RF : Random Forest

RNN : Recurrent Neural Network

ROC: Receiver Operating Characteristic

SVM: Support Vector Machine

XGB: XGBoost (Extreme Gradient Boosting)

INTRODUCTION GÉNÉRALE

La douleur abdominale aiguë constitue l'un des motifs de consultation les plus fréquents dans les services d'urgence, représentant environ 5 à 10 % de l'ensemble des admissions non traumatiques [1,3]. Ce symptôme recouvre un spectre étiologique remarquablement large, depuis des pathologies bénignes et spontanément résolutive, telles que les gastro-entérites infectieuses, jusqu'à des urgences chirurgicales majeures engageant le pronostic vital à court terme, comme la perforation d'ulcère, l'ischémie mésentérique ou la péritonite généralisée [5,8]. Le caractère souvent trompeur de la présentation clinique initiale, la diversité des organes potentiellement impliqués et la subjectivité de la sensation douloureuse rendent le diagnostic étiologique particulièrement difficile, en particulier chez les sujets âgés ou immunodéprimés [1,7].

L'approche diagnostique conventionnelle repose sur une anamnèse minutieuse, un examen physique systématisé et des examens complémentaires biologiques et d'imagerie [1,4]. Cependant, la performance de cette démarche est limitée. Les scores clinico-biologiques classiques, tels que le score d'Alvarado ou l'*Appendicitis Inflammatory Response* (AIR), ont été développés pour standardiser la suspicion diagnostique, principalement dans l'appendicite aiguë [11,23]. Si le score AIR offre des performances satisfaisantes, avec une aire sous la courbe proche de 0,89 [12], ces outils restent spécifiques à une pathologie unique et ne couvrent pas l'ensemble des causes chirurgicales possibles. De plus, leur précision diminue dans certaines populations, comme les enfants, les femmes enceintes ou les patients très âgés [7,11]. Ainsi, la décision de recourir à une intervention chirurgicale ou, au contraire, d'opter pour une abstention et une surveillance demeure fréquemment complexe et source d'erreurs, qu'il s'agisse de laparotomies blanches ou de retards de prise en charge préjudiciables [4,17].

Dans ce contexte, l'intelligence artificielle (IA) a émergé comme une voie prometteuse pour améliorer la précision diagnostique et la stratification du risque dans les situations d'urgence [24,25]. Les systèmes d'aide à la décision clinique (*Clinical Decision Support Systems*, CDSS) basés sur l'IA exploitent des algorithmes d'apprentissage automatique et profond pour analyser de manière simultanée un grand nombre de variables cliniques, biologiques et radiologiques [28,35]. Dans le domaine de la douleur abdominale aiguë, plusieurs modèles de *machine learning* ont montré des performances élevées, avec des aires sous la courbe dépassant 0,90 pour prédire la nécessité d'une intervention chirurgicale [32,39]. Toutefois, la majorité de ces modèles n'ont pas été validés dans des conditions cliniques réelles, leur explicabilité reste souvent insuffisante, et leur généralisabilité à des populations variées n'est pas garantie [27,42,43].

L'intégration des données cliniques (anamnèse, examen physique, paramètres vitaux) et biologiques (numération formule sanguine, marqueurs inflammatoires, bilan hépatique et pancréatique) dans des modèles prédictifs basés sur l'IA pourrait révolutionner l'approche diagnostique de la douleur abdominale aiguë [9,17,18]. Ces systèmes d'aide à la décision clinique (Clinical Decision Support Systems - CDSS) pourraient assister les cliniciens dans leur raisonnement diagnostique, réduire les délais de prise en charge et améliorer les outcomes des patients [13].

C'est dans cette optique que s'inscrit le présent travail. L'objectif principal de cette étude était d'évaluer la performance diagnostique d'un système d'IA pré-entraîné, Khashef AI (basé sur DEEP SEEK V4) , pour l'identification du risque d'urgence chirurgicale chez des patients consultant pour douleur abdominale aiguë dans deux centres hospitaliers algériens. Nous avons mené une étude transversale de validation d'un test diagnostique, en comparant les prédictions du modèle (niveau de triage) à un *gold standard* clinique et paraclinique établi indépendamment. L'analyse a porté sur un échantillon de 87 patients pour lesquels le statut d'urgence chirurgicale était documenté de manière certaine. Les indicateurs de performance calculés — sensibilité, spécificité, valeurs prédictives, rapports de vraisemblance et précision globale — ont permis d'apprécier la capacité du modèle à exclure ou à confirmer une urgence, et de discuter sa place potentielle dans la stratégie de triage aux urgences.

Objectifs de la recherche:

Objectif principale :

- Valider une application pratique et ergonomique intégrant le modèle (DEEP SEEK V4) performant.

Les objectifs secondaires

- Valider le système développé sur une cohorte indépendante de patients
- Évaluer la sensibilité , la spécificité , VPN et VPP du modèle

PREMIÈRE PARTIE

CADRE THÉORIQUE

CHAPITRE 1

LA DOULEUR ABDOMINALE AIGÜE

1.1 Définition et concepts fondamentaux

La douleur abdominale aiguë se définit comme une sensation douloureuse non traumatique, d'apparition récente (moins de 7 jours), localisée dans la région comprise entre le diaphragme et le plancher pelvien [3]. Elle constitue un motif majeur de recours aux services d'urgence, représentant 5 à 10 % de toutes les consultations dans ces structures [4,1]. La difficulté diagnostique réside dans la diversité des organes potentiellement impliqués et dans le caractère subjectif de la sensation douloureuse, qui peut être d'origine viscérale, somatique ou projetée [2]. La reconnaissance précoce d'une urgence chirurgicale est un enjeu vital car un retard de prise en charge aggrave le pronostic [1].

.

1.1.1 Mécanismes physiopathologiques de la douleur abdominale

Trois grands types de mécanismes sont classiquement distingués :

- **Douleur viscérale vraie** : elle résulte de la stimulation des nocicepteurs présents dans la paroi des viscères creux (intestin, voies biliaires, uretères) ou de la capsule des organes pleins (foie, rate). La distension brutale d'un organe creux, l'ischémie ou la contraction violente du muscle

lisse activent des fibres afférentes non myélinisées de type C qui cheminent avec les nerfs sympathiques. La sensation est typiquement sourde, mal localisée, médiane, et s'accompagne fréquemment de manifestations végétatives (nausées, sueurs) [9,2].

- **Douleur somatique (pariétale)** : produite par l'irritation du péritoine pariétal, elle est transmise par des fibres A- δ myélinisées. Elle est précise, latéralisée, exacerbée par les mouvements et s'accompagne d'une contracture réflexe de la paroi abdominale. L'inflammation d'origine appendiculaire, cholécystitique ou pancréatique évolue souvent d'une douleur viscérale initiale vers une douleur somatique lorsque l'inflammation atteint le péritoine pariétal [1].

- **Douleur projetée (référée)** : la convergence des afférences viscérales et somatiques sur un même segment médullaire peut faire ressentir la douleur dans un territoire cutané éloigné de l'organe lésé. L'exemple classique est la douleur scapulaire droite de la cholécystite, en rapport avec une irritation du nerf phrénique (C4) [2].

La compréhension de ces mécanismes est essentielle pour interpréter correctement la sémilogie et éviter les erreurs d'orientation diagnostique [1].

1.2 Classification de la douleur abdominale

Plusieurs systèmes de classification ont été proposés pour organiser l'approche diagnostique de la douleur abdominale aiguë. Les plus utilisés sont la classification topographique et la classification étiologique.

1.2.1 Classification topographique

La répartition en quadrants ou en régions abdominales est la méthode la plus utilisée en pratique courante :

- **Douleur de l'hypochondre droit** : évoque une pathologie hépatobiliaire (cholécystite, colique hépatique), une hépatite aiguë, un abcès hépatique ou une pathologie du rein droit (pyélonéphrite, colique néphrétique) [1].

- **Douleur épigastrique** : oriente vers une pathologie gastroduodénale (ulcère perforé, gastrite), une pancréatite aiguë, ou un infarctus du myocarde à expression abdominale [1].

- **Douleur de l'hypochondre gauche** : suggère une splénomégalie douloureuse, un infarctus splénique, une pathologie du rein gauche ou une colite gauche [6].

- **Douleur péri-ombilicale** : typique de l'appendicite débutante (stade viscéral) ou d'une occlusion intestinale grêle [4].

- **Douleur de la fosse iliaque droite** : caractérise l'appendicite aiguë au stade somatique, une pathologie ovarienne droite, une iléite terminale (maladie de Crohn) ou une colique néphrétique [1].

- **Douleur de la fosse iliaque gauche** : évoque une sigmoïdite diverticulaire, une colite, une pathologie ovarienne gauche ou une colique néphrétique [1].

- **Douleur diffuse** : peut correspondre à une péritonite généralisée, une pancréatite aiguë grave, une occlusion intestinale étendue ou une urgence vasculaire (ischémie mésentérique) [5].

Cette classification topographique, bien que très utile, doit toujours être confrontée à l'examen physique et aux examens complémentaires, car la localisation peut être trompeuse à la phase précoce de nombreuses affections [6].

Tableau 1.1: Classification des douleurs abdominales selon la localisation

Région	Organes concernés	Pathologies fréquentes
Épigastre	Estomac, duodénum, pancréas	Ulcère, gastrite, pancréatite
Hypochondre droit	Foie, vésicule biliaire	Cholécystite, colique hépatique, hépatite
Hypochondre gauche	Rate, angle colique gauche	Splénomégalie, infarctus splénique
Région péri-ombilicale	Intestin grêle, aorte	Appendicite (début), occlusion, anévrisme
Fosse iliaque droite	Appendice, annexes droites	Appendicite, pathologie annexielle
Fosse iliaque gauche	Sigmoïde, annexes gauches	Sigmoïdite, pathologie annexielle
Hypogastre	Vessie, utérus, rectum	Cystite, pathologie utérine
Flancs	Reins, uretères, côlon	Colique néphrétique, pyélonéphrite

1.2.2 Classification étiologique

D'un point de vue pratique, il est utile de distinguer les causes chirurgicales, nécessitant une intervention rapide, des causes médicales pouvant bénéficier d'un traitement conservateur. Cette distinction guide la prise en charge initiale et l'orientation du patient.

1.3 Épidémiologie

La douleur abdominale aiguë représente un problème de santé publique majeur à l'échelle mondiale, avec des implications significatives en termes de morbidité, de mortalité et de coûts de santé.

1.3.1 Données statistiques

La douleur abdominale aiguë est une cause extrêmement fréquente de consultation. Aux États-Unis, elle est responsable d'environ 7 millions de visites annuelles aux services d'urgence [10]. En Europe, une étude prospective multicentrique a rapporté que 8 à 10 % des admissions non traumatiques aux urgences étaient motivées par une douleur abdominale [5]. La prévalence est comparable dans les pays à revenu intermédiaire, bien que le profil étiologique diffère (infections parasitaires, tuberculose abdominale) [8].

L'âge moyen des patients se situe autour de 45 ans, avec une légère prédominance féminine (sex-ratio de 1,2 :1) [3]. La douleur abdominale est le premier motif de consultation pour les patients de plus de 65 ans, chez qui la mortalité liée aux urgences abdominales est trois fois plus élevée que dans la population générale, en raison du retard diagnostique et des comorbidités [7].

1.3.2 Facteurs de risque

Plusieurs facteurs augmentent la probabilité d'une cause chirurgicale sous-jacente :

- **Âge avancé** : au-delà de 65 ans, le risque de trouver une étiologie chirurgicale (appendicite compliquée, perforation digestive, ischémie mésentérique) est significativement plus élevé [7].

- **Sexe** : les femmes présentent davantage de pathologies pelviennes (grossesse extra-utérine, torsion d'annexe), tandis que les hommes sont plus exposés aux ulcères perforés et aux pancréatites alcooliques [1].
- **Antécédents chirurgicaux abdominaux** : les brides post-opératoires sont la première cause d'occlusion du grêle [5].
- **Immunodépression (VIH, corticothérapie, chimiothérapie)** : favorise les infections opportunistes et les perforations sur lymphome digestif [1].
- **Comorbidités cardiovasculaires** : fibrillation auriculaire, athérosclérose augmentent le risque d'ischémie mésentérique aiguë [8].
- **Prise d'AINS ou de corticoïdes** : masque la symptomatologie et favorise les perforations digestives [4].

1.4 Étiologies principales

Les étiologies de la douleur abdominale aiguë sont nombreuses et variées. Une approche systématique distinguant les causes chirurgicales des causes médicales facilite le raisonnement diagnostique.

1.4.1 Causes chirurgicales urgentes

Les urgences chirurgicales abdominales les plus fréquentes, validées par une vaste revue systématique [5], sont :

- **Appendicite aiguë** : première cause d'abdomen aigu chirurgical chez l'adulte jeune (20-30 ans). Le diagnostic repose sur la douleur de la fosse iliaque droite, la défense et les signes biologiques inflammatoires. La tomodensitométrie abdominale a amélioré la précision diagnostique [1].
- **Cholécystite aiguë** : deuxième cause, souvent précédée de coliques hépatiques. L'échographie abdominale montre un épaississement pariétal vésiculaire > 4 mm et un signe de Murphy échographique [1].

- **Occlusion intestinale aiguë** : le plus souvent sur brides, parfois sur tumeur colique ou volvulus. La distension abdominale, les vomissements et l'arrêt des matières et des gaz dominent le tableau. Le scanner permet d'identifier le niveau et la cause de l'obstruction [7].
- **Perforation d'ulcère gastroduodénal** : douleur épigastrique intense, en « coup de poignard », contracture abdominale généralisée. Le pneumopéritoine visualisé sur une radiographie d'abdomen sans préparation ou au scanner confirme le diagnostic [8].
- **Ischémie mésentérique aiguë** : urgence vasculaire grave avec mortalité élevée (50-80 %). Douleur disproportionnée par rapport à l'examen physique, acidose métabolique et élévation des lactates. L'angioscanner mésentérique est l'examen de référence [4].
- **Pathologies pelviennes** : grossesse extra-utérine rompue, torsion d'annexe, salpingite aiguë. Le dosage des β -HCG et l'échographie endovaginale sont indispensables [1].
- **Autres urgences** : pancréatite aiguë nécrosante, colique néphrétique compliquée (pyonéphrose), anévrisme de l'aorte abdominale fissuré (douleur lombaire irradiant vers l'abdomen) [3].

1.4.2 Causes médicales

Un grand nombre de douleurs abdominales aiguës relèvent de pathologies médicales ne nécessitant pas de chirurgie immédiate :

- **Gastro-entérite aiguë infectieuse** : diarrhée et vomissements prédominant. La coproculture et la PCR virale permettent l'identification de l'agent (norovirus, rotavirus, salmonellose) [5].
- **Coliques hépatiques et néphrétiques simples** : douleur paroxystique sans signe infectieux ni retentissement rénal. Traitement antalgique et expulsion spontanée du calcul [1].
- **Pathologies inflammatoires intestinales** : maladie de Crohn et rectocolite hémorragique en poussée. Le scanner abdomino-pelvien et l'iléocoloscopie affinent le diagnostic [1].
- **Syndrome de l'intestin irritable (SII)** : trouble fonctionnel fréquent dont les exacerbations peuvent simuler un abdomen aigu. L'absence de signes objectifs d'inflammation et la chronicité des symptômes orientent [6].
- **Causes métaboliques et toxiques** : acidocétose diabétique (douleur diffuse, vomissements), crise de porphyrie aiguë intermittente, intoxication au plomb, hypercalcémie [4].
- **Pathologies thoraciques à expression abdominale** : infarctus du myocarde inférieur, pneumonie de la base droite, embolie pulmonaire. L'ECG et la radiographie thoracique sont systématiques devant une douleur épigastrique isolée [3].

- **Causes pariétales** : hématome de la gaine du muscle droit, hernie étranglée spontanément réductible, zona avant éruption [1].

1.5 Approche diagnostique conventionnelle

L'approche diagnostique de la douleur abdominale aiguë repose sur une démarche clinique rigoureuse et structurée, combinant l'anamnèse, l'examen physique et les examens complémentaires [2]

1.5.1 Anamnèse

L'interrogatoire demeure l'étape pivot de l'orientation. Il doit préciser :

- **Mode d'installation** : brutal (perforation, colique) ou progressif (appendicite, cholécystite).
- **Siège et irradiations** : déplacements évocateurs (douleur épigastrique migrant en fosse iliaque droite pour l'appendicite).
- **Type de douleur** : crampe (obstruction), brûlure (ulcère), torsion (torsion d'annexe), continue (péritonite).
- **Facteurs déclenchants ou calmants** : alimentation, position, médicaments (AINS).
- **Signes d'accompagnement** : fièvre, vomissements, arrêt des gaz et des selles, diarrhée, signes urinaires, toux, dyspnée.
- **Antécédents personnels et chirurgicaux** : gestes abdominaux antérieurs, maladies chroniques, traitement en cours (anticoagulants, corticoïdes).
- **Chez la femme en âge de procréer** : date des dernières règles, notion de retard, contraception [1,3].

La qualité de l'anamnèse permet une première hypothèse diagnostique dans près de 70 % des cas [4].

1.5.2 Examen physique

L'examen clinique systématisé complète l'anamnèse et recherche les signes objectifs de gravité :

- **Inspection** : distension, cicatrices, mouvements respiratoires abdominaux (diminués en cas de péritonite).
- **Auscultation** : silence auscultatoire (iléus paralytique ou péritonite), borborygmes métalliques (occlusion mécanique).
- **Palpation** : recherche d'une défense (contracture réflexe localisée) ou d'une contracture généralisée (péritonite diffuse). La palpation douce permet de repérer une masse (abcès, anévrisme). La douleur à la décompression (signe de Blumberg) traduit une irritation péritonéale.
- **Percussion** : disparition de la matité pré-hépatique (pneumopéritoine), douleur à la percussion (signe de Murphy pour la cholécystite).
- **Touchers pelviens** : le toucher rectal et le toucher vaginal recherchent un bombement du cul-de-sac de Douglas (péritonite, hémopéritoine), une masse pelvienne ou une douleur latéralisée.
- **Évaluation hémodynamique** : tension artérielle, fréquence cardiaque, signes de choc (marbrures, oligurie) qui imposent une réanimation concomitante à la démarche diagnostique [3].

L'examen physique doit être répété à intervalles réguliers car un tableau initial rassurant peut rapidement se dégrader, notamment chez les sujets âgés ou immunodéprimés [7].

CHAPITRE 2

PARAMÈTRES CLINIQUES ET BIOLOGIQUES

2.1 Examen clinique

L'examen clinique reste le premier temps de l'évaluation d'une douleur abdominale aiguë. Il associe le recueil des paramètres vitaux et la recherche de signes physiques spécifiques, dont la valeur diagnostique a été largement étudiée [1].

2.1.1 Paramètres vitaux

La mesure systématique de la fréquence cardiaque, de la pression artérielle, de la température corporelle et de la fréquence respiratoire permet d'identifier rapidement les patients à risque de dégradation. La tachycardie (fréquence cardiaque > 100 battements/min) est souvent le premier signe de réponse systémique à une agression intra-abdominale et constitue un indicateur indépendant de gravité, en particulier lorsqu'elle est associée à une hypotension [17]. Une hypotension artérielle (pression systolique < 90 mmHg) traduit un choc septique ou hypovolémique et impose une réanimation immédiate et la recherche d'une urgence chirurgicale (péritonite, ischémie mésentérique) [3]. La fièvre ($> 38^{\circ}\text{C}$) est inconstante mais sa présence renforce la suspicion d'une étiologie infectieuse ou inflammatoire (cholécystite, appendicite) ; toutefois, l'absence de fièvre n'élimine pas une infection grave, notamment chez le sujet âgé [15]. La polypnée, quant à elle, peut témoigner d'une acidose métabolique (ischémie mésentérique, pancréatite aiguë sévère) ou d'une irritation diaphragmatique [1].

2.1.2 Signes physiques spécifiques

La palpation abdominale recherche des signes d'irritation péritonéale : défense, contracture, douleur à la décompression (signe de Blumberg). Ces signes possèdent une spécificité élevée pour le diagnostic de péritonite, bien que leur sensibilité soit modérée, surtout chez les patients âgés ou immunodéprimés [1].

Plusieurs manœuvres ciblent des pathologies particulières. Le **signe de Murphy** (arrêt de l'inspiration profonde lors de la palpation de l'hypochondre droit) présente une sensibilité de 65-85 % et une spécificité de 85-95 % pour la cholécystite aiguë [16].

Dans l'appendicite aiguë, le **signe du psoas** (douleur à l'extension de la cuisse droite) et le **signe de l'obturateur** (douleur à la rotation interne de la cuisse fléchie) sont évocateurs d'une localisation rétro-cæcale ou pelvienne de l'appendice, mais leur sensibilité isolée reste faible [1,11]. La présence d'une douleur à la palpation en fosse iliaque droite, combinée à une défense, améliore la probabilité d'appendicite, en particulier chez l'adulte jeune [11].

Le **signe de Rovsing** (douleur en fosse iliaque droite lors de la palpation de la fosse iliaque gauche) possède une spécificité rapportée de 75-90 % pour l'appendicite, bien que sa sensibilité ne dépasse pas 30-50 % [1]. L'examen clinique doit également comporter la recherche d'une masse abdominale battante (anévrisme de l'aorte abdominale), la percussion des fosses lombaires (colique néphrétique) et l'inspection des orifices herniaires [15].

L'association de plusieurs signes cliniques améliore la précision diagnostique, mais aucun signe isolé n'est suffisant pour affirmer ou exclure une urgence chirurgicale ; c'est pourquoi les scores clinico-biologiques ont été développés [11].

Tableau 2.1: Signes cliniques spécifiques et pathologies évoquées [1]

Signe	Technique	Pathologie évoquée	Sensibilité
Signe de Murphy	Douleur à l'inspiration lors de la palpation d	e l'hypochondre droit Cholécystite aiguë	65-95%
Signe de McBurney	Douleur au point de McBurney (jonction 1/3	ext - 2/3 int ligne ombiliAppendicite	c-EIAS) 50-94%
Signe de Blumberg	Douleur à la décompression brutale de l'ab	domen Irritation péritonéale	80-90%
Signe de Rovsing	Douleur FID lors de la palpation de la FIG	Appendicite	22-68%
Signe du psoas	Douleur à l'extension passive de la hanche	droite Appendicite rétrocæca	le13-42%
Signe de l'obturateur	Douleur à la rotation interne de la hanche fl	échie Appendicite pelvienne	8-25%
Défense abdominale	Contraction réflexe de la paroi à la palpation	n Irritation péritonéale	Variable
Contracture	Rigidité permanente "ventre de bois"	Péritonite	Très spécifique

2.2 Paramètres biologiques

Les examens biologiques complètent l'évaluation clinique en apportant des données objectives sur l'état inflammatoire, infectieux et métabolique du patient. Leur interprétation doit toujours être corrélée au contexte clinique.

2.2.1 Numération Formule Sanguine (NFS)

L'hyperleucocytose ($> 10\,000/\text{mm}^3$) et la neutrophilie sont les anomalies les plus fréquentes au cours des syndromes abdominaux aigus inflammatoires. Une méta-analyse récente portant sur l'appendicite aiguë a montré une sensibilité de 78 % (IC 95 % : 74-82 %) et une spécificité de 67 % (IC 95 % : 62-72 %) pour la leucocytose seule [19]. La présence d'une déviation gauche (augmentation des polynucléaires neutrophiles immatures) renforce l'hypothèse infectieuse, tandis que la combinaison d'une leucocytose normale et d'une CRP normale possède une excellente valeur prédictive négative [19].

Le rapport neutrophiles/lymphocytes (NLR) a été proposé comme marqueur de sévérité dans l'appendicite et la cholécystite. Une méta-analyse a retrouvé une sensibilité de 86 % et une spécificité de 71 % pour le diagnostic d'appendicite compliquée lorsque le NLR était supérieur à 4,7 [18].

2.2.2 Marqueurs inflammatoires

La protéine C-réactive (CRP) est synthétisée par le foie en réponse à l'interleukine-6. Sa concentration augmente dans les 6 à 12 heures suivant le début d'un processus inflammatoire et atteint un pic à 48 heures. Dans l'appendicite aiguë, la CRP possède une spécificité de 75-80 %, mais une sensibilité plus faible dans les premières heures (40-60 %), ce qui limite son intérêt isolé au stade précoce [19]. La combinaison CRP-leucocytose améliore la performance diagnostique [11].

La procalcitonine (PCT) est un marqueur plus spécifique d'infection bactérienne systémique. Une méta-analyse a suggéré qu'une valeur de PCT $> 0,5$ ng/mL pouvait discriminer une appendicite compliquée (perforée ou gangréneuse) d'une appendicite simple avec une sensibilité

de 62 % et une spécificité de 75 % [23]. Toutefois, la PCT n'est pas recommandée en routine dans l'évaluation initiale de toute douleur abdominale, mais peut aider au diagnostic de sepsis intra-abdominal [11].

2.2.3 Bilan hépatique et pancréatique

Le dosage de la lipase sérique est le marqueur de référence pour le diagnostic de pancréatite aiguë : une élévation supérieure à trois fois la normale supérieure, associée à une douleur épigastrique typique, confirme le diagnostic avec une sensibilité de 85-95 % et une spécificité > 95 % [20].

En présence d'une douleur de l'hypochondre droit, le bilan hépatique (transaminases, phosphatases alcalines, bilirubine) recherche une obstruction biliaire ou une hépatite aiguë. Une élévation isolée des ALAT (> 100 UI/L) oriente vers une hépatite, tandis qu'une hyperbilirubinémie conjuguée et une élévation des phosphatases alcalines suggèrent une lithiase de la voie biliaire principale [21].

2.2.4 Autres paramètres biologiques

- **Lactates sérique** : une hyperlactatémie (> 2 mmol/L) témoigne d'une hypoperfusion tissulaire et constitue un facteur de mauvais pronostic en cas d'ischémie mésentérique, de pancréatite sévère ou de sepsis. Une revue systématique a montré une valeur prédictive élevée des lactates pour la mortalité à 30 jours dans les urgences abdominales [22].

- **Créatinine et urée** : l'insuffisance rénale aiguë fonctionnelle accompagne souvent les états septiques graves ou la déshydratation secondaire aux vomissements.

- **Bandelette urinaire et β -HCG** : la recherche d'une hématurie oriente vers une colique néphrétique ; chez la femme en âge de procréer, le dosage des β -HCG est indispensable pour exclure une grossesse extra-utérine [1,15].

La confrontation des données biologiques à la clinique permet d'affiner la probabilité diagnostique et de guider la prescription des examens d'imagerie [11].

2.3 Scores diagnostiques existants

Face à la complexité du diagnostic de certaines pathologies abdominales, plusieurs scores clinico-biologiques ont été développés pour standardiser l'évaluation et améliorer la

précision diagnostique. Ces scores combinent des paramètres cliniques et biologiques pour calculer une probabilité diagnostique.

2.3.1 Score d'Alvarado

Le score d'Alvarado (ou MANTRELS), décrit en 1986, attribue des points à huit items : migration de la douleur, anorexie, nausées/vomissements, défense en fosse iliaque droite, douleur à la décompression, élévation de la température, hyperleucocytose et déviation gauche. Le score total varie de 0 à 10. Un score < 4 exclut l'appendicite avec une sensibilité proche de 96 % ; un score > 7 la confirme avec une spécificité d'environ 80-85 % [23]. Cependant, une méta-analyse récente a montré que le score d'Alvarado présente une aire sous la courbe ROC de 0,82, ce qui le rend moins performant chez l'enfant, la femme enceinte et le sujet âgé. Sa principale faiblesse réside dans la subjectivité de certains items (migration de la douleur, anorexie) [11,23].

Tableau 2.3: Score d'Alvarado pour l'appendicite

Critère	Points
M - Migration de la douleur vers la FID	1
A - Anorexie	1
N - Nausées/Vomissements	1
T - Tenderness (Douleur) en FID	2
R - Rebound (Défense/Blumberg)	1
E - Élévation de la température (> 37.3°C)	1
L - Leucocytose (> 10 000/mm ³)	2
S - Shift (Neutrophilie > 75%)	1
TOTAL	10

2.3.2 Score AIR (Appendicitis Inflammatory Response)

Le score AIR, proposé par Andersson en 2008, repose sur des variables objectives : vomissements, douleur en fosse iliaque droite, défense ou douleur à la décompression, température > 38,5°C, leucocytose, pourcentage de polynucléaires neutrophiles et CRP. Le score s'échelonne de 0 à 12. Une étude de validation externe a rapporté une aire sous la courbe ROC de 0,89, significativement supérieure à celle du score d'Alvarado (0,81) [12,13]. Le score AIR permet une stratification plus fine du risque : un score < 4 exclut l'appendicite avec une sensibilité de 97 % et une valeur prédictive négative de 96 %, tandis qu'un score > 8 prédit une appendicite avec une spécificité de 95 % [13]. Les recommandations de la World Society of Emergency Surgery (WSES) 2020 le considèrent comme le score de première intention chez l'adulte [11].

2.3.3 Autres scores

Plusieurs autres scores ont été proposés pour améliorer la précision diagnostique ou pour s'adapter à des populations particulières :

- **Score RIPASA (Raja Isteri Pengiran Anak Saleha Appendicitis)** : développé en Malaisie, il inclut des paramètres démographiques, cliniques et biologiques. Une revue systématique a montré une sensibilité de 98 % mais une spécificité plus faible (45-50 %) que le score AIR, ce qui peut conduire à un excès d'imagerie [14].
- **Score PAS (Pediatric Appendicitis Score)** : adapté à l'enfant, il intègre la douleur à la toux, la défense et l'absence de signes urinaires. Sa performance est comparable à celle du score d'Alvarado en pédiatrie [11].
- **Score AAS (Adult Appendicitis Score)** : propose une stratification en trois niveaux (risque faible, intermédiaire, élevé) et intègre des paramètres cliniques et biologiques. Sa capacité à réduire le recours à la tomodensitométrie sans perte de sécurité diagnostique est en cours d'évaluation [14].
- **Comparaison des scores**: une méta-analyse a conclu que le score AIR offre le meilleur compromis entre sensibilité et spécificité chez l'adulte, suivi par le score AAS, tandis que le

score d'Alvarado conserve un intérêt en contexte de ressources limitées en raison de sa simplicité [13,11].

CHAPITRE 3

L'INTELLIGENCE ARTIFICIELLE EN MÉDECINE

3.1 Fondements de l'intelligence artificielle

L'intelligence artificielle (IA) désigne la capacité d'un système informatique à exécuter des tâches qui requièrent habituellement l'intelligence humaine, comme le raisonnement, l'apprentissage, la perception ou la prise de décision [24]. Dans le domaine médical, l'IA a connu un essor remarquable grâce à l'augmentation exponentielle des données de santé numérisées et aux progrès des algorithmes d'apprentissage automatique [25].

3.1.1 Définition et concepts de base

L'IA s'appuie sur deux branches principales : l'apprentissage automatique (**machine learning, ML**), qui permet à un algorithme d'identifier des motifs dans les données sans être explicitement programmé pour chaque règle, et l'apprentissage profond (**deep learning, DL**), une sous-catégorie du ML utilisant des réseaux de neurones artificiels à multiples couches [24,26]. Les réseaux de neurones convolutifs (CNN) excellent dans l'analyse d'images tandis que les réseaux récurrents (RNN) et les transformeurs sont adaptés aux données séquentielles ou textuelles [26].

L'entraînement d'un modèle d'IA nécessite un jeu de données annoté (apprentissage supervisé) ou non (apprentissage non supervisé). La validation externe sur des populations indépendantes est indispensable pour garantir la généralisabilité des performances [25]. La qualité et la représentativité des données conditionnent l'absence de biais algorithmiques [27].

3.1.2 Types d'intelligence artificielle

On distingue classiquement :

- **IA étroite (faible)** : spécialisée dans une tâche unique (ex. classification d'images radiologiques). C'est la quasi-totalité des applications actuelles en médecine [24].
- **IA générale (forte)** : hypothétique, capable de résoudre tout problème intellectuel humain. Elle n'est pas encore réalisée [25].

En médecine, l'IA est majoritairement une IA étroite, déployée comme outil d'aide et non comme substitut au clinicien. Les systèmes sont conçus pour augmenter les capacités diagnostiques, standardiser les pratiques et réduire la charge cognitive [28].

3.2 Applications médicales de l'IA

Les applications de l'IA en santé couvrent désormais un large spectre, de la radiologie à la médecine interne, en passant par la dermatologie et l'ophtalmologie [25]. Dans le contexte de la douleur abdominale aiguë, deux domaines sont particulièrement pertinents : l'imagerie médicale et le diagnostic fondé sur les données clinico-biologiques.

3.2.1 Imagerie médicale

L'analyse automatisée d'images radiologiques a constitué le terrain de développement le plus précoce et le plus fructueux de l'IA médicale. Des études multicentriques ont démontré que des algorithmes de deep learning peuvent identifier une appendicite aiguë sur une tomodensitométrie (TDM) abdominale avec une aire sous la courbe (AUC) dépassant 0,90, soit des performances proches de celles de radiologues expérimentés [29]. Un modèle basé sur CNN a atteint une sensibilité de 93 % et une spécificité de 91 % pour le diagnostic d'appendicite, tout en réduisant le temps d'interprétation [29].

D'autres travaux ont porté sur la détection de la cholécystite aiguë par échographie. Un système combinant segmentation automatique et classification a montré une précision de 92 % pour le diagnostic de cholécystite lithiasique [30]. L'IA peut également aider à la quantification de l'infiltration graisseuse péri-pancréatique pour évaluer la sévérité d'une pancréatite aiguë [31].

Ces outils ne se substituent pas au radiologue mais peuvent servir de « seconde lecture » ou orienter les examens urgents en priorisant les dossiers suspects [29].

3.2.2 Diagnostic basé sur les données cliniques

Les données structurées (âge, sexe, paramètres vitaux, résultats biologiques) et non structurées (texte des comptes rendus, antécédents) peuvent être exploitées par des modèles de ML pour prédire une étiologie chirurgicale ou un risque de complication [28]. Une revue systématique a analysé 18 études utilisant des algorithmes de ML (forêts aléatoires, XGBoost, réseaux de neurones) pour le diagnostic de l'appendicite aiguë. L'AUC moyenne était de 0,85 à 0,95, dépassant significativement les scores cliniques traditionnels comme le score d'Alvarado [32].

Dans une étude prospective portant sur 725 patients admis pour douleur abdominale aiguë, un modèle de gradient boosting a identifié les patients nécessitant une intervention chirurgicale avec une sensibilité de 91 % et une spécificité de 83 %, surpassant les cliniciens juniors [33]. L'intégration de variables cliniques simples (leucocytose, CRP, localisation de la douleur) dans un algorithme permet une stratification fiable du risque et réduit potentiellement le recours inutile au scanner [33,34].

Les modèles de traitement du langage naturel (NLP) commencent aussi à être utilisés pour extraire automatiquement des informations pertinentes des dossiers médicaux électroniques, facilitant l'identification précoce des urgences abdominales [28].

3.3 Systèmes d'aide à la décision clinique (CDSS)

Les systèmes d'aide à la décision clinique (**Clinical Decision Support Systems, CDSS**) sont des logiciels qui fournissent au clinicien des informations et des recommandations personnalisées au moment de la prise en charge du patient [35]. Les CDSS basés sur l'IA intègrent des algorithmes de ML entraînés sur des données massives et produisent des prédictions ou des alertes en temps réel [35,36]

3.3.1 Composantes d'un CDSS

Un CDSS classique repose sur trois composantes [35] :

1. **Base de connaissances** : contient les règles médicales, les données cliniques de référence et les modèles prédictifs. Dans les CDSS basés sur l'IA, cette base est constituée du modèle entraîné et de ses poids.
2. **Moteur d'inférence** : applique les règles ou le modèle aux données spécifiques du patient pour générer une hypothèse ou un score.
3. **Interface utilisateur** : intégrée au dossier médical électronique ou sous forme de plateforme web, elle permet au clinicien de visualiser la recommandation et le niveau de confiance associé.

Les CDSS modernes peuvent utiliser des données en temps réel (signes vitaux, résultats biologiques) pour générer des alertes précoces, par exemple en cas de sepsis intra-abdominal [36].

3.3.2 Avantages et défis des CDSS basés sur l'IA

Avantages [13]:

- **Amélioration de la précision diagnostique** : plusieurs études ont montré que les CDSS réduisent le taux d'erreurs diagnostiques dans les services d'urgence [34]. Un système basé sur un réseau bayésien a permis de diminuer de 18 % les admissions inutiles pour douleur abdominale non spécifique [37].
- **Standardisation des pratiques** : l'IA applique des critères homogènes, indépendamment de la fatigue ou de l'expérience du clinicien [28].
- **Gain de temps** : l'analyse automatisée des données accélère la prise de décision, réduisant le délai avant intervention chirurgicale dans l'appendicite aiguë [33].

Défis [14,15] :

- **Qualité des données** : la performance du modèle dépend de la qualité et de l'exhaustivité des données d'entraînement. Des biais de sélection peuvent entraîner des erreurs systématiques chez certaines populations (minorités ethniques, sujets âgés) [27].
- **Explicabilité** : de nombreux algorithmes de deep learning fonctionnent comme des « boîtes noires », rendant difficile la compréhension de leur raisonnement par le clinicien, ce qui freine la confiance et l'adoption [25,38].
- **Intégration clinique** : l'implémentation d'un CDSS nécessite une interopérabilité avec les systèmes d'information hospitaliers, une formation des personnels et une validation prospective de l'impact clinique [35].
- **Responsabilité médico-légale** : la question de la responsabilité en cas d'erreur de diagnostic imputable à un système d'IA n'est pas entièrement résolue sur le plan juridique [36].

Malgré ces défis, les CDSS basés sur l'IA représentent une avancée prometteuse pour améliorer la prise en charge des douleurs abdominales aiguës, en complément de l'expertise clinique [34,38].

3.4 Architecture et conception des grands modèles de langage (LLMs) pour le diagnostic médical

3.4.1 Définition et principes fondamentaux des grands modèles de langage

Les grands modèles de langage, ou *Large Language Models* (LLMs), constituent une catégorie de modèles d'apprentissage profond entraînés sur des corpus textuels massifs afin de comprendre, de générer et de raisonner sur le langage naturel. Leur émergence a marqué une rupture technologique dans le domaine de l'intelligence artificielle, en permettant le traitement de tâches linguistiques d'une complexité sans précédent, allant de la traduction automatique à la rédaction de textes techniques, en passant par le raisonnement structuré.

Dans le domaine médical, les LLMs ont démontré un potentiel transformateur, notamment pour l'analyse de textes cliniques, le soutien au diagnostic, la génération de comptes rendus et l'aide à la décision. Une revue systématique récente couvrant la période 2022–2025 a confirmé que GPT-4 et ses variantes dominent la recherche (74 % des études recensées), suivis par GPT-3.5 (36 %), avec des applications majeures dans la classification diagnostique, la réponse à des questions médicales et la génération de contenu clinique [26].

La particularité des LLMs réside dans leur **capacité d'apprentissage auto-supervisé** : au lieu d'être explicitement programmés pour une tâche spécifique, ils apprennent les régularités statistiques du langage en prédisant le mot suivant dans une séquence textuelle (apprentissage auto-régressif) ou en reconstituant des segments masqués (apprentissage par masquage). Cette approche leur confère une plasticité remarquable, leur permettant de s'adapter à une multitude de tâches médicales sans nécessiter une reprogrammation Complète [26, 27].

3.4.2 Architecture fondamentale : le modèle Transformer

Le mécanisme d'attention : rupture conceptuelle

L'architecture à la base de tous les LLMs modernes est le **Transformer**, introduit en 2017 dans l'article fondateur *Attention Is All You Need*. Avant cette innovation, les modèles de traitement du langage reposaient principalement sur les réseaux de neurones récurrents (RNNs) et leurs variantes (LSTMs, GRUs), qui traitaient le texte de manière séquentielle, mot par mot. Cette approche souffrait de limitations majeures : traitement lent et difficulté à capturer les dépendances à longue distance au sein d'un texte.

Le Transformer a révolutionné le domaine en remplaçant la récurrence par un **mécanisme d'attention**, qui permet au modèle de pondérer simultanément l'importance relative de chaque élément d'une séquence par rapport à tous les autres. Mathématiquement, le mécanisme d'attention calcule des pondérations reflétant la pertinence de chaque partie de l'entrée pour la tâche à accomplir, puis applique ces pondérations pour moduler l'influence de chaque segment. Ce mécanisme s'inspire directement de la cognition humaine : face à une phrase complexe, le lecteur accorde sélectivement plus d'attention aux mots saillants et ignore les détails secondaires [28].

3.4.3 Les trois familles architecturales

L'architecture Transformer se décline en trois configurations principales, chacune optimisée pour des usages distincts :

L'architecture encodeur seul (Encoder-only) : Représentée par BERT (Bidirectional Encoder Representations from Transformers) et ses dérivés médicaux (ClinicalBERT, BioBERT). Cette architecture traite le langage de manière bidirectionnelle (de gauche à droite et de droite à gauche simultanément), ce qui lui confère une compréhension contextuelle profonde. Elle excelle dans les tâches de classification, d'extraction d'information et de découverte de connaissances médicales [27]. .

L'architecture décodeur seul (Decoder-only) : Représentée par les modèles GPT (dont GPT-4) et DeepSeek V4. Cette architecture traite le langage de manière unidirectionnelle (auto-régressive), générant le texte mot après mot. Elle est idéale pour les tâches de génération de contenu, d'interaction avec les patients et de raisonnement diagnostique structuré. Les modèles de cette famille sont ceux ayant démontré les meilleures performances dans le diagnostic médical, avec une amélioration continue des capacités de raisonnement clinique [29, 30].

L'architecture décodeur seul (Decoder-only) : Représentée par les modèles GPT (dont GPT-4) et DeepSeek V4. Cette architecture traite le langage de manière unidirectionnelle (auto-régressive), générant le texte mot après mot. Elle est idéale pour les tâches de génération de contenu, d'interaction avec les patients et de raisonnement diagnostique structuré. Les modèles de cette famille sont ceux ayant démontré les meilleures performances dans le diagnostic médical, avec une amélioration continue des capacités de raisonnement clinique.

Pour l'application visée dans notre étude — le diagnostic différentiel de la douleur abdominale aiguë — l'architecture décodeur seul (GPT, DeepSeek) est la plus pertinente, car elle permet de générer un raisonnement clinique structuré et d'émettre des hypothèses diagnostiques ordonnées par probabilité, mimant ainsi le processus cognitif du clinicien [30].

3.4.4 Innovations architecturales avancées

L'architecture Mixture-of-Experts (MoE)

L'une des innovations majeures adoptées par les LLMs de dernière génération, dont DeepSeek V4, est l'architecture dite de **Mélange d'Experts (Mixture-of-Experts, MoE)**. Contrairement aux modèles denses traditionnels où tous les paramètres sont sollicités pour chaque entrée, l'architecture MoE segmente le modèle en multiples sous-modules spécialisés (les "experts"). Pour chaque requête, un mécanisme de routage sélectionne dynamiquement un sous-ensemble d'experts pertinents, n'activant ainsi qu'une fraction des paramètres totaux.

Cette conception présente des avantages décisifs pour le domaine médical :

- **Efficacité computationnelle** : un modèle MoE peut atteindre une capacité de raisonnement très élevée (plusieurs centaines de milliards de paramètres totaux) tout en maintenant un coût d'inférence contenu, car seuls les paramètres "actifs" sont sollicités. DeepSeek-R1, par exemple, totalise 671 milliards de paramètres mais n'en active qu'une fraction lors de l'inférence [29].
- **Spécialisation médicale** : des travaux récents, comme Med-MoE et SparseDoctor, ont démontré la pertinence de l'architecture MoE pour le domaine médical : différents experts peuvent se spécialiser implicitement dans différentes sous-disciplines (gastro-entérologie, urologie, gynécologie), améliorant la précision diagnostique sans compromettre la généralité du modèle [26].

Le raisonnement par chaîne de pensée (Chain-of-Thought)

Les LLMs les plus avancés intègrent désormais des capacités de raisonnement pas-à-pas, dites *Chain-of-Thought (CoT)*. Cette technique permet au modèle de décomposer un problème complexe — comme un diagnostic différentiel — en étapes intermédiaires explicites, améliorant significativement la qualité et la transparence du raisonnement clinique final.

Dans le contexte de notre étude, cette capacité est fondamentale : face à un tableau de douleur abdominale aiguë, DeepSeek V4 peut analyser séquentiellement les données démographiques, les antécédents, la sémilogie douloureuse, les signes cliniques, les résultats biologiques et les données d'imagerie, puis synthétiser ces informations en une hypothèse diagnostique cohérente avec ses probabilités associées. Cette approche structurée se rapproche du raisonnement clinique humain et facilite la validation et la vérification des prédictions par le praticien [30]. .

CHAPITRE 4

ÉTAT DE L'ART

4.1 Modèles IA existants pour la douleur abdominale

L'application de l'intelligence artificielle au diagnostic et au pronostic des douleurs abdominales aiguës a fait l'objet d'un nombre croissant de travaux depuis 2018. La majorité des modèles publiés visent à différencier les urgences chirurgicales des causes médicales, à prédire la nécessité d'une intervention ou à améliorer le diagnostic précoce de l'appendicite aiguë [39,40].

4.1.1 Études récentes majeures

Plusieurs études ont démontré la supériorité des algorithmes d'apprentissage automatique sur les scores cliniques classiques. Une revue systématique publiée en 2022, portant sur 29 études, a montré que les modèles de machine learning atteignaient une AUC moyenne de 0,89 (IC 95 % : 0,86-0,92) pour le diagnostic d'appendicite aiguë, contre 0,78 pour le score d'Alvarado [32]. Les algorithmes de type **gradient boosting** (XGBoost) et forêts aléatoires présentaient les meilleures performances [32].

Une étude rétrospective multicentrique incluant 11 000 patients adultes a développé un modèle de réseau de neurones pour prédire le caractère chirurgical d'une douleur abdominale. L'AUC était de 0,91 en validation externe, avec une sensibilité de 92 % et une spécificité de 84 % [39]. L'ajout des données de l'examen clinique et des antécédents améliorait significativement la performance par rapport aux modèles fondés uniquement sur les données biologiques [39].

Concernant la population pédiatrique, un modèle de deep learning entraîné sur 5 000 dossiers a montré une capacité à réduire de 27 % le recours à la tomodensitométrie sans augmenter le taux d'appendicectomies blanches [40]. L'algorithme utilisait des variables cliniques, des résultats biologiques et des scores de douleur.

En imagerie, l'utilisation de CNN pour l'analyse automatisée du scanner abdominal a démontré des performances diagnostiques comparables à celles de radiologues seniors pour la détection de l'appendicite et de la cholécystite [29,30]. Un système hybride combinant analyse d'image et données textuelles du dossier médical a permis d'améliorer le diagnostic d'ischémie mésentérique, avec une sensibilité de 88 % contre 72 % pour l'interprétation humaine seule [41].

Tableau 4.1: Études récentes sur l'IA et la douleur abdominale

Auteur (Année)	Pathologie	Modèle	N	AUC/Accuracy
Park et al. (2024) [58]	Appendicite	XGBoost	2,847	AUC: 0.91
Chen et al. (2024) [59]	DAA multiple	Random Forest	1,523	Acc: 85.3%
Kim et al. (2024) [11]	Appendicite péd.	Deep Learning	1,089	AUC: 0.89
Marcinkevics (2024) [64]	Appendicite	XGBoost	782	AUC: 0.87
Akmese (2024) [60]	Appendicite	Ensemble ML	1,234	Acc: 94%
Wang et al. (2024) [65]	DAA triage	Random Forest	3,421	AUC: 0.85

4.1.2 Variables d'entrée utilisées

Les modèles exploitent principalement des données structurées et non structurées. Les variables les plus fréquemment intégrées sont :

- **Données démographiques** : âge, sexe [39,32].
- **Paramètres vitaux** : fréquence cardiaque, pression artérielle, température, saturation en oxygène [33,34].
- **Caractéristiques cliniques** : localisation et type de douleur, signes péritonéaux (défense, contracture), signes spécifiques (Murphy, psoas), durée des symptômes [11,33].
- **Biologie standard** : leucocytose, neutrophilie, CRP, rapport neutrophiles/lymphocytes, lipase, bilan hépatique [19,32,39].
- **Données d'imagerie** : images brutes ou rapports structurés de scanner, échographie [29,30,41].

- **Antécédents médicaux** : comorbidités cardiovasculaires, antécédents chirurgicaux abdominaux, traitements en cours [34,39].

Certains modèles récents incluent des scores textuels extraits par traitement automatique du langage à partir des comptes rendus cliniques [28,41]. La multimodalité, combinant texte et images, est de plus en plus explorée pour améliorer la robustesse prédictive [40].

4.2 Limites des approches actuelles et perspectives

4.2.1 Limites identifiées

Plusieurs freins limitent encore l'adoption clinique des modèles d'IA pour la douleur abdominale aiguë :

L'analyse critique de la littérature révèle plusieurs limites des approches actuelles:
Généralisation limitée: La plupart des modèles sont développés et validés sur des populations spécifiques (un seul centre, une population ethnique particulière). Leur performance peut se dégrader significativement lors de l'application à d'autres contextes [24].

Focus sur une pathologie unique: La majorité des études se concentrent sur l'appendicite, négligeant le diagnostic différentiel global de la douleur abdominale qui inclut de multiples étiologies [10].

Manque d'interprétabilité: Les modèles de type "boîte noire" (notamment les réseaux de neurones profonds) sont difficiles à interpréter, ce qui limite leur acceptation par les cliniciens et pose des questions de responsabilité médicale [19,25].

Qualité des données: Les études rétrospectives sont soumises à des biais de sélection et à des données manquantes. La standardisation des données d'entrée est souvent insuffisante [24].

Validation externe insuffisante: Peu d'études incluent une validation sur une cohorte externe indépendante, limitant la confiance dans la généralisabilité des résultats [24,13].

Intégration clinique: Très peu de systèmes ont été effectivement déployés et évalués en conditions réelles d'utilisation clinique [13,14].

4.2.2 Perspectives et positionnement de notre recherche

- **Qualité et hétérogénéité des données** : la plupart des études sont rétrospectives et monocentriques. Les données utilisées sont souvent incomplètes ou codées de façon non standardisée, ce qui affecte la généralisabilité [27,42].
- **Risque de biais algorithmique** : les populations d'entraînement sous-représentent fréquemment les minorités ethniques, les sujets très âgés ou les patients atteints de pathologies atypiques [27]. Cela peut entraîner des erreurs systématiques lorsqu'un modèle est déployé dans un environnement différent.
- **Manque de validation externe prospective** : seule une minorité d'études testent leur modèle dans un cadre clinique réel ou comparent l'impact sur la décision médicale [34,43]. L'absence de validation prospective rend difficile l'évaluation du bénéfice net.
- **Explicabilité insuffisante** : les modèles de type « boîte noire » (deep learning) ne fournissent pas d'explication compréhensible au clinicien, ce qui nuit à la confiance et peut masquer des erreurs de raisonnement [25,38].
- **Intégration aux systèmes d'information hospitaliers** : les contraintes techniques et réglementaires (interopérabilité, sécurité des données, certification en tant que dispositif médical) freinent l'implémentation à grande échelle [35,42].

DEUXIÈME PARTIE

ÉTUDE PRATIQUE

CHAPITRE 5 :

MATÉRIEL ET MÉTHODES

5.1 Type et conception de l'étude

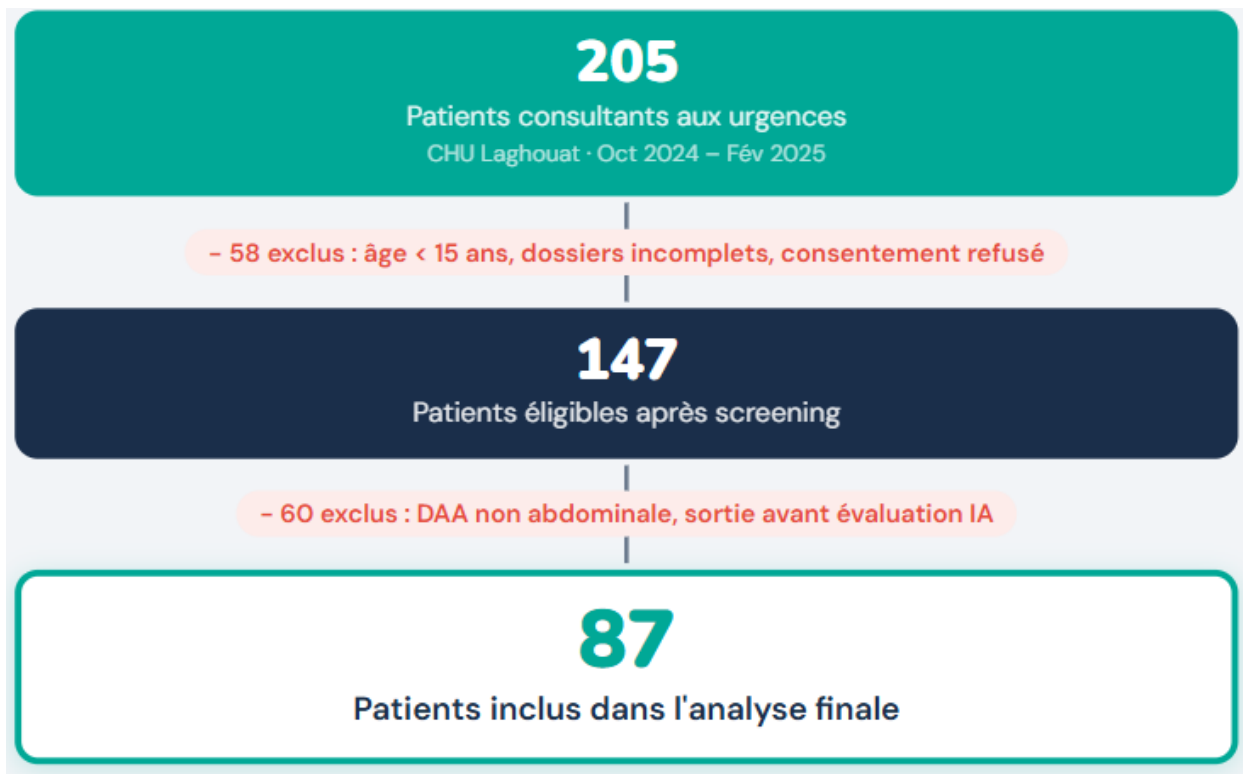
Il s'agit d'une étude transversale, de validation d'un test diagnostique. L'objectif principal est d'évaluer la performance d'un système d'IA pré-entraîné, Kashef AI (basé sur DEEP SEEK V4), dans l'identification du risque d'urgence chirurgicale chez des patients se présentant pour douleur abdominale aiguë.

La collecte initiale des données, effectuée aux niveau de pavillon d'urgence de l'hôpital mixte Laghouat et service d'urgence chirurgicale de l'hôpital Mustafa-Bacha , a alimenté une base de données brute qui a ensuite été soumise au modèle d'IA pour classification. Les prédictions du modèle ont été comparées à un étalon-or (gold standard) clinique et paraclinique afin de déterminer ses propriétés intrinsèques (sensibilité, spécificité) et extrinsèques (valeurs prédictives) .

5.2 Population et échantillonnage

5.2.1 Population source

La population étudiée est composée de patients adultes et adolescents consultant pour une douleur abdominale aiguë. Les données initiales, collectées dans le tableau, regroupent 205 dossiers patients (numérotés de 01 à 205). Chaque entrée correspond à un patient unique, documenté par un identifiant, un sexe, un âge, ses antécédents, les caractéristiques sémiologiques de sa douleur, les résultats de sa biologie et de son imagerie.



5.2.2 Critères de sélection

Pour la phase d'analyse de performance, seuls les dossiers pour lesquels le diagnostic de référence (gold standard) a pu être clairement établi ont été retenus, conformément à la méthodologie de validation. L'analyse statistique finale a ainsi porté sur un sous-ensemble de **87 patients** pour lesquels la présence ou l'absence d'une urgence chirurgicale était documentée de manière certaine.

Les critères d'inclusion pour cette analyse étaient :

- Patients pour lesquels le diagnostic final était clairement établi (confirmation chirurgicale, imagerie caractéristique, ou évolution clinique documentée)
- Dossiers complets permettant une classification binaire (urgence chirurgicale : oui/non)

Les critères d'exclusion comprenaient les patients perdus de vue, les dossiers incomplets, et les cas où le diagnostic de référence restait incertain.

5.3 Variables de l'étude et collecte des données

5.3.1 Variables d'entrée (Features) pour le modèle d'IA

Le modèle **Kashef Ai** a reçu de manière exhaustive l'ensemble des données cliniques, biologiques et d'imagerie disponibles pour chaque patient. Selon le fichier d'analyse du modèle, les variables soumises au système incluaient .

Données démographiques :

- Sexe (Masculin/Féminin)
- Âge

Paramètres cliniques :

- Caractéristiques de la douleur : localisation, type (brutale/progressive), caractère (brûlure, compression, crampe, coup de couteau), durée, intensité (sur 10)
- Irradiation de la douleur
- Signes spécifiques : signe de Murphy, signe de McBurney, signe de Blumberg, signe de Rovsing
- Symptômes associés : nausées, vomissements, diarrhée, constipation, fièvre, arrêt des matières et des gaz
- Antécédents médicaux et chirurgicaux
- Toucher rectal (TR)

Paramètres biologiques :

- Numération Formule Sanguine : leucocytes (GB), hémoglobine (HGB), plaquettes (PLT)
- Protéine C-Réactive (CRP)
- Ionogramme sanguin : natrémie (Na⁺), kaliémie (K⁺)
- Bilan hépatique : ASAT, ALAT, bilirubine totale
- Lipasémie
- Créatininémie, urée

Cette approche exhaustive est cohérente avec les recommandations de la littérature qui identifient ces paramètres comme les plus discriminants pour le diagnostic différentiel de la douleur abdominale aiguë. L'analyse de l'importance des variables dans les études antérieures a confirmé le rôle central des marqueurs inflammatoires (CRP, leucocytes) et des signes cliniques spécifiques .

Figure 6.1 : Interface et module de localisation de la douleur

KaShef Ai

Assistant de Diagnostic Médical pour la Douleur Abdominale

Avertissement Médical Important : Cette application est destinée à des fins éducatives et de guidance uniquement. Elle ne remplace pas une consultation médicale professionnelle. Consultez toujours un médecin pour un diagnostic précis.

Note Technique : Cette application utilise une clé API pour le traitement des diagnostics. Tous les calculs sont effectués via notre service cloud sécurisé.

Bienvenue dans KaShef Ai

Cette application vous aidera à explorer les diagnostics différentiels possibles pour la douleur abdominale.

Commencer le Diagnostic

Hypochondre G	Épigastre	Hypochondre D
Région G	Ombilic	Région D
Illaque G	Hypogastre	Illaque D

Région sélectionnée: Illaque droit

Figure 6.2 : Module d'informations cliniques

Informations Cliniques

Veuillez fournir les informations cliniques pertinentes pour le diagnostic.

Informations du Patient

Âge:

Entrez l'âge du patient en années

Sexe:

Sélectionnez le sexe du patient

Caractéristiques de la Douleur

Début de la douleur:

Indiquez comment la douleur a commencé

Durée:

Indiquez la durée de la douleur

Intensité (0-10): 5

Début de la douleur:

Indiquez comment la douleur a commencé

Durée:

Indiquez la durée de la douleur

Intensité (0-10): 5

Déplacez le curseur pour indiquer l'intensité de la douleur (0 = pas de douleur, 10 = douleur insoutenable)

Caractère de la douleur:

Décrivez le caractère de la douleur

Facteurs aggravants:

Facteurs qui aggravent la douleur

Facteurs sédatifs:

Facteurs qui soulagent la douleur

Symptômes Associés

☐ Nausées ☐ Vomissements ☐ Fièvre ☐ Diarrhée ☐ Constipation ☐ Ballonnements

Figure 6.3 : Module antécédents et données biologiques

Antécédents

Médicaux:

Hypertension
Diabète
Maladie cardiaque
Maladie rénale
Maladie hépatique
Maladie inflammatoire intestinale

Sélectionnez les antécédents médicaux (maintenez Ctrl pour en sélectionner plusieurs)

Chirurgicaux:

Appendicectomie
Cholécystectomie
Réparation de hernie
Chirurgie intestinale
Chirurgie gynécologique
Aucun

Sélectionnez les antécédents chirurgicaux (maintenez Ctrl pour en sélectionner plusieurs)

Données Biologiques

Leucocytes ($\times 10^9/L$):

Valeur normale: 4.0-11.0

CRP (mg/L):

Valeur normale: < 5.0

Hémoglobine (g/dL):

Plaquettes ($\times 10^9/L$):

Valeur normale: 150-450

Sodium (mmol/L):

Valeur normale: 135-145

Potassium (mmol/L):

Valeur normale: 3.5-5.0

Créatinine ($\mu\text{mol/L}$):

Valeur normale: 50-120

Bilirubine ($\mu\text{mol/L}$):

Valeur normale: < 20

Retour

Traiter le Diagnostic

Résultats du Diagnostic

Résumé du Cas

Région sélectionnée: Iliaque droit

Âge: ans

Sexe:

Caractéristiques de la douleur: (5/10), , début

Facteurs aggravants: Aucun

Facteurs sédatifs: Aucun

Symptômes associés: Aucun

Antécédents médicaux: Aucun

Antécédents chirurgicaux: Aucun

Données biologiques: Aucune

Diagnostic Différentiel (Classé par Probabilité)

1. Appendicite

40%

Preuves cliniques: Douleur dans le quadrant inférieur droit avec signes inflammatoires

Explication: Inflammation de l'appendice provoquant une douleur caractéristique dans le quadrant inférieur droit.

2. Maladie de Crohn

20%

Preuves cliniques: Douleur iliaque droite chronique avec diarrhée

Explication: Maladie inflammatoire chronique de l'intestin pouvant affecter n'importe quelle partie du tube digestif.

3. Ovaire (torsion ou kyste)

15%

Preuves cliniques: Douleur iliaque droite chez une femme avec irrégularités menstruelles

Explication: Pathologies ovariennes pouvant causer une douleur dans le quadrant inférieur droit chez les femmes.

Imprimer les Résultats

Exporter les Résultats

Sauvegarder le Cas

Nouveau Diagnostic

5.3.2 Variables de sortie (Outputs) du modèle

Pour chaque patient, le modèle Kashef AI a généré trois types de prédictions structurées :

1. **Niveau de triage (Triage) :** Un code couleur et alphanumérique évaluant le degré de gravité et l'urgence de la prise en charge :
 - **N1 :** Extrême urgence
 - **N2 :** Urgence
 - **N3 :** Urgence relative / Non urgent
 - **N4 :** Pas d'urgence
2. **Diagnostic principal :** Le diagnostic étiologique le plus probable, estimé par le modèle, accompagné de sa probabilité en pourcentage (entre parenthèses).
3. **Diagnostic secondaire :** Un diagnostic différentiel alternatif, avec sa probabilité estimée.
4. **Signaux d'alarme :** Paramètres cliniques ou biologiques anormaux justifiant une attention particulière.
5. **Examens suggérés :** Examens complémentaires recommandés par le modèle.

5.4 Développement et évaluation du modèle

5.4.1 Architecture du modèle Kashef A

Le modèle kashef AI est un système d'IA pré-entraîné, dont l'architecture interne et les paramètres d'apprentissage sont considérés comme une "boîte noire" dans le cadre de cette validation clinique. L'évaluation a porté exclusivement sur la performance diagnostique finale de ses prédictions, sans accès aux mécanismes internes de décision.

Cette approche d'évaluation externe, sans connaissance des paramètres d'entraînement ni de l'architecture exacte du modèle, reflète une situation réaliste de déploiement clinique où les praticiens interagissent avec des systèmes d'IA pré-entraînés sans nécessairement comprendre leur fonctionnement interne .

5.4.2 Binarisation du test diagnostique

À des fins d'évaluation quantitative, le niveau de triage qualitatif a été converti en un test diagnostique binaire :

- Les patients classés en niveaux de triage "**N1**" (**Extrême urgence**) ou "**N2**" (**Urgence**) ont été considérés comme ayant un **test positif**, c'est-à-dire que le modèle signale un risque élevé d'urgence chirurgicale.
- Les patients classés "**N3**" (**Urgence relative/Non urgent**) ont été considérés comme ayant un **test négatif**.
- Aucun patient n'a été classé N4 dans la cohorte analysée .

5.4.3 Gold standard (vérité terrain)

La vérité terrain (gold standard) a été établie indépendamment pour chaque patient en se basant sur :

- Les comptes rendus opératoires pour les cas chirurgicaux
- Les résultats d'imagerie interprétés par des radiologues seniors
- L'évolution clinique documentée
- Les conclusions explicites de prise en charge (urgence chirurgicale, absence d'urgence, orientation vers d'autres spécialités)

Ces informations ont été consignées dans les fichiers sources .

5.5 Analyse statistique

L'analyse de la performance diagnostique a été réalisée par la construction d'un **tableau de contingence 2×2** (matrice de confusion) comparant le résultat du test (N1/N2 vs N3) au diagnostic de référence (Urgence chirurgicale : oui/non) .

Les indicateurs de performance suivants ont été calculés :

Indicateurs intrinsèques (indépendants de la prévalence) :

- **Sensibilité (Se)** : Capacité du modèle à identifier correctement les patients nécessitant une intervention chirurgicale urgente.
 - Formule : $VP / (VP + FN)$

- **Spécificité (Sp)** : Capacité du modèle à exclure correctement les cas non urgents
 - Formule : $VN / (VN + FP)$

Indicateurs extrinsèques (dépendants de la prévalence) :

- **Valeur Prédictive Positive (VPP)** : Probabilité qu'un patient classé N1/N2 soit réellement une urgence chirurgicale
 - Formule : $VP / (VP + FP)$

- **Valeur Prédictive Négative (VPN)** : Probabilité qu'un patient classé N3 ne soit pas une urgence chirurgicale
 - Formule : $VN / (VN + FN)$

Indicateurs synthétiques :

- **Précision globale (Accuracy)** : Proportion de patients correctement classés

- Formule : $(VP + VN) / \text{Total}$

- **Prévalence (P)** : Proportion réelle d'urgences chirurgicales dans l'échantillon

- Formule : $(VP + FN) / \text{Total}$

- **Rapport de Vraisemblance Positif (LR+)** : Combien de fois un test positif est plus probable chez un patient nécessitant une urgence que chez un patient n'en nécessitant pas

- Formule : $Se / (1 - Sp)$

- **Rapport de Vraisemblance Négatif (LR-)** : Combien de fois un test négatif est plus probable chez un patient sans urgence que chez un patient avec urgence

- Formule : $(1 - Se) / Sp$

L'ensemble de ces analyses a été réalisé sur un échantillon de **87 patients validés**, pour lesquels les données de référence étaient complètes et fiables .

L'analyse statistique adoptée suit les recommandations internationales pour l'évaluation des modèles prédictifs en médecine, notamment le cadre TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis), qui préconise une validation rigoureuse des modèles avec des métriques standardisées.

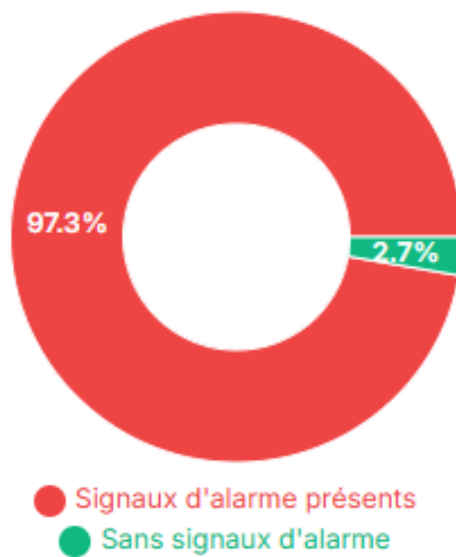
CHAPITRE 6 :

RÉSULTATS.

6.1 Caractéristiques de la population et classifications du modèle

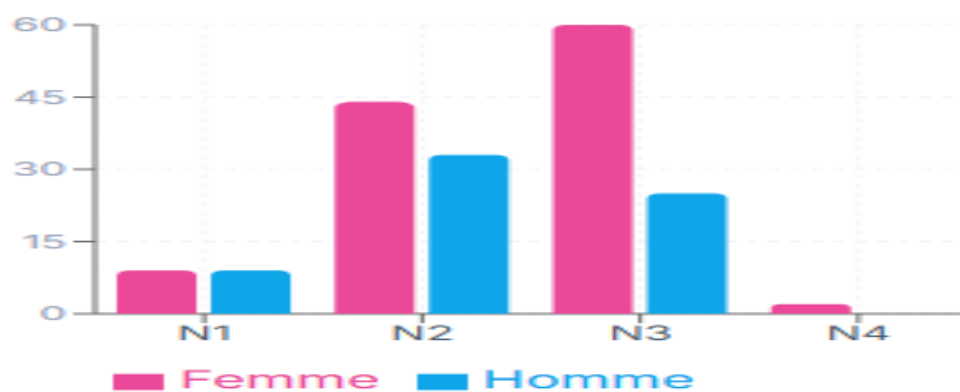
Signaux d'Alarme

Présence de signaux cliniques d'alarme



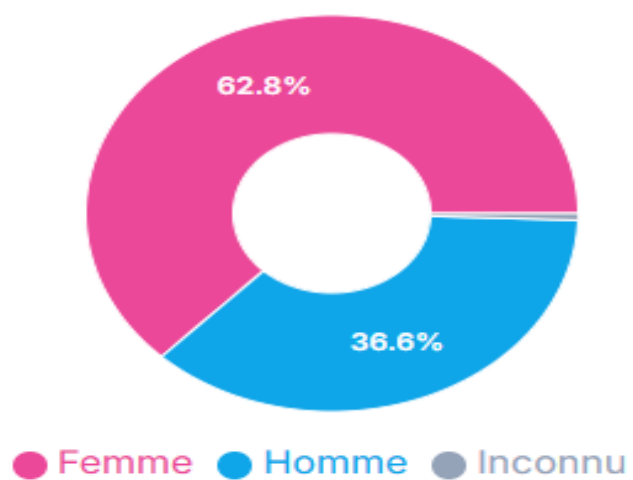
Triage par Sexe

Comparaison hommes/femmes par niveau d'urgence



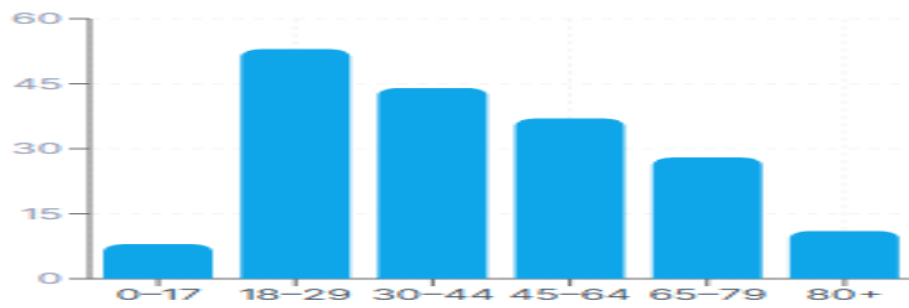
Répartition par Sexe

Proportion hommes / femmes



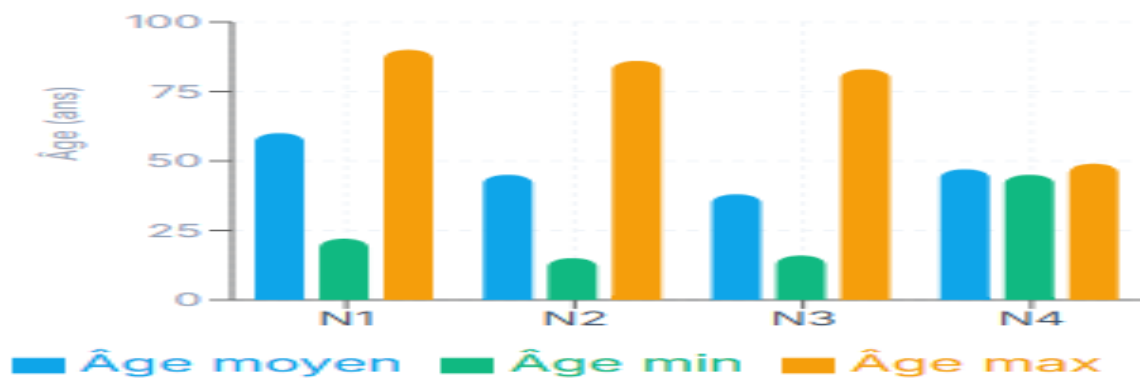
Distribution des Tranches d'Âge

Nombre de patients par groupe d'âge



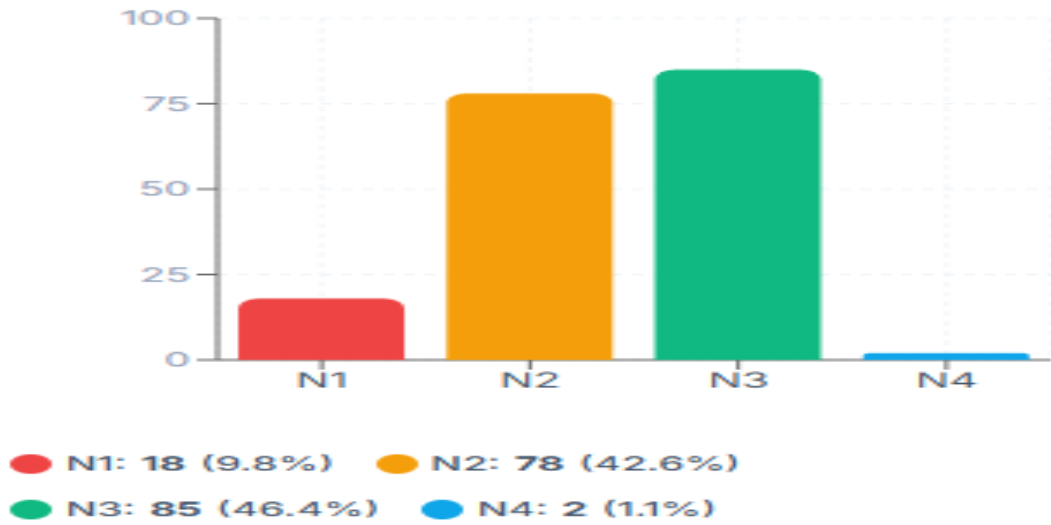
Âge Moyen par Niveau de Triage

Statistiques d'âge par niveau d'urgence



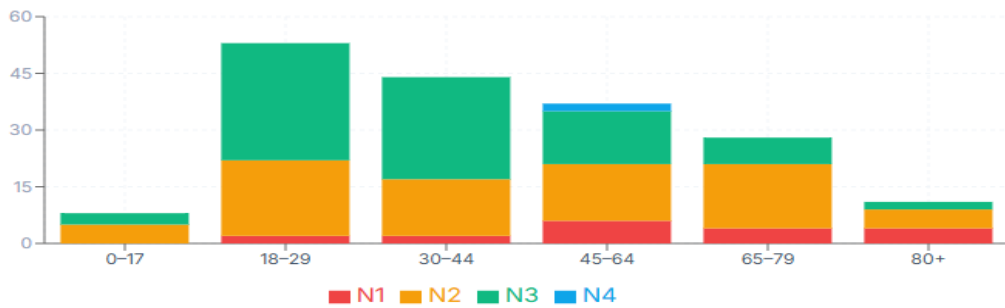
Répartition par Niveau de Triage

Distribution des patients selon l'urgence



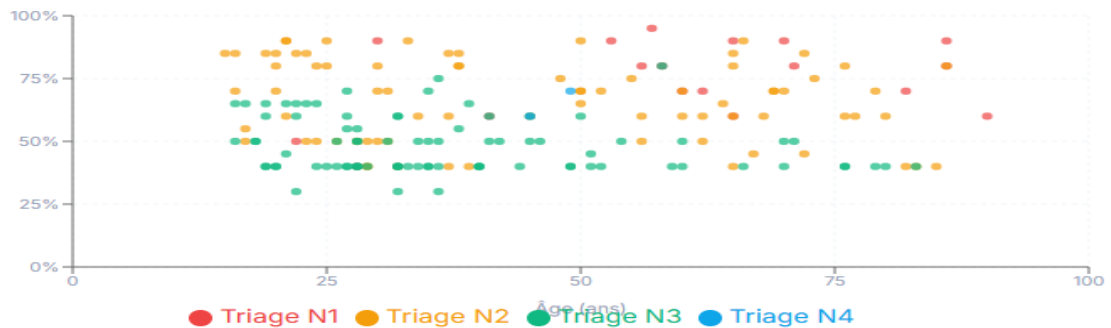
Triage par Tranche d'Âge

Répartition de l'urgence selon l'âge



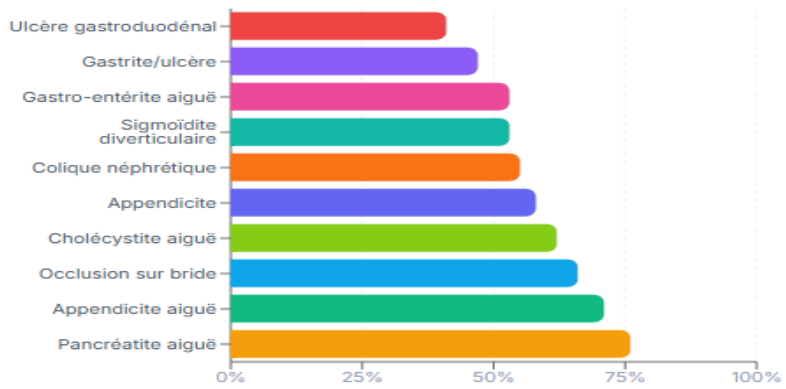
Âge vs Probabilité du Diagnostic Principal

Corrélation entre l'âge et la certitude diagnostique par triage



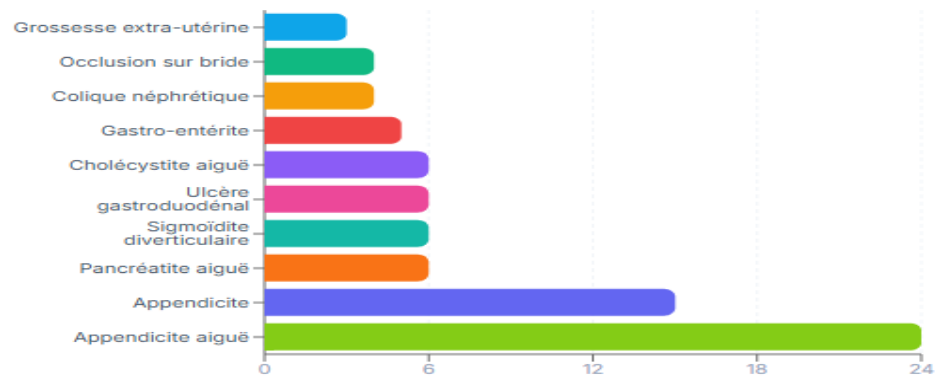
Probabilité Moyenne des Diagnostics Principaux

Diagnostics avec ≥ 3 cas, triés par certitude moyenne



Top 10 Diagnostics Principaux

Diagnostics les plus fréquents



6.1.1 Distribution des niveaux de triage

Le Tableau 6.1 détaille les prédictions pour l'ensemble des 87 dossiers soumis au modèle. L'analyse de la répartition des niveaux de triage, tous diagnostics confondus, révèle une distribution spécifique des classifications de gravité.

La répartition des niveaux de triage est la suivante :

Tableau 6.1 : L'analyse de la répartition des niveaux de triage

Niveau de Triage	Effectif	Pourcentage
N1 (Extrême urgence)	8	9,2%
N2 (Urgence)	37	42,5%
N3 (Urgence relative/Non urgent)	41	47,1%
N4 (Pas d'urgence)	1	1,2%
Total	87	100%

Le niveau de triage le plus fréquemment attribué est le niveau N3 (urgence relative / non urgent), représentant 41 des 87 patients de la cohorte (47,1%). Le niveau N2 (urgence) a été attribué à 37 patients (42,5%). Une minorité de patients a été classée en niveau N1 (extrême urgence), soit 8 patients (9,2%). Seul 1 patient a été classé N4 (1,2%).

En combinant les niveaux de triage maximaux, un total de 45 patients sur l'ensemble de la cohorte a été signalé comme étant à risque d'urgence (N1 ou N2), tandis que 42 patients ont été classés comme non urgents (N3 ou N4).

6.1.2 Diagnostics principaux générés par le modèle

Le modèle Kashef AI a généré une grande variété de diagnostics principaux, couvrant les principales étiologies de la douleur abdominale aiguë. Les diagnostics les plus fréquemment évoqués comme diagnostic principal ou secondaire incluent :

- Appendicite aiguë
- Occlusion intestinale (sur bride, sténose, fécalome)
- Péritonite (sur perforation, hernie étranglée)
- Pancréatite aiguë
- Cholécystite aiguë
- Angiocholite aiguë
- Colique néphrétique
- Pathologies gynécologiques (torsion d'annexe, kyste ovarien, grossesse extra-utérine)
- Pathologies non chirurgicales (gastro-entérite, colopathie fonctionnelle, infection urinaire)

Cette diversité diagnostique reflète le spectre étiologique large de la douleur abdominale aiguë, tel que décrit dans le cadre théorique, allant de pathologies bénignes auto-résolutives à des urgences chirurgicales mettant en jeu le pronostic vital .

6.2 Performance diagnostique pour l'identification de l'urgence chirurgicale

L'évaluation de la validité du modèle a été conduite sur le sous-ensemble de 87 patients pour lesquels le statut d'urgence chirurgicale (oui/non) a pu être déterminé avec .

6.2.1 Prévalence de l'urgence chirurgicale

Dans cet échantillon analysé, la prévalence réelle de l'urgence chirurgicale était de **18,4 %** (16 patients sur 87). La majorité des patients (71 sur 87, soit 81,6%) ne relevait pas d'une prise en charge chirurgicale urgente.

Cette prévalence relativement faible est cohérente avec les données épidémiologiques de la littérature. Les études rapportent que parmi les patients consultant pour douleur abdominale aiguë aux urgences, 20 à 30% présentent une pathologie chirurgicale, l'appendicite aiguë représentant la cause la plus fréquente d'abdomen aigu chirurgical (25 à 30% des urgences chirurgicales abdominales) .

6.2.2 Matrice de confusion

Tableau 6.2 : La matrice de confusion 2×2, comparant le test de triage binaire (N1/N2 = Positif, N3 = Négatif) au diagnostic de référence

	Référence Positive (Urgence chirurgicale)	Référence Négative (Pas d'urgence)	Total
Modèle Positif (N1/N2)	Vrais Positifs (VP) = 12	Faux Positifs (FP) = 35	47
Modèle Négatif (N3)	Faux Négatifs (FN) = 4	Vrais Négatifs (VN) = 36	40
Total	16	71	87

Analyse détaillée de la matrice :

- **Vrais Positifs (VP) = 12** : Le modèle a correctement identifié 12 patients comme étant à risque d'urgence chirurgicale, et ces patients étaient effectivement des urgences chirurgicales.
- **Faux Positifs (FP) = 35** : Le modèle a classé 35 patients en N1/N2, suggérant une urgence chirurgicale, alors que ces patients ne relevaient pas d'une prise en charge chirurgicale urgente.
- **Faux Négatifs (FN) = 4** : Le modèle a classé 4 patients en N3, suggérant l'absence d'urgence, alors que ces patients nécessitaient une intervention chirurgicale urgente.
- **Vrais Négatifs (VN) = 36** : Le modèle a correctement classé 36 patients comme non urgents, et ces patients ne relevaient effectivement pas d'une urgence chirurgicale.

6.2.3 Indicateurs de performance calculés

À partir de cette matrice de confusion, les paramètres statistiques suivants ont été **Indicateurs intrinsèques** :

1. **Sensibilité (Se)** : $12 / (12 + 4) = 12 / 16 = 75,0 \%$

Le modèle a correctement identifié 3 patients sur 4 nécessitant une intervention chirurgicale urgente. Un quart des urgences chirurgicales (4 patients sur 16) n'a pas été détecté par le système.

2. **Spécificité (Sp)** : $36 / (36 + 35) = 36 / 71 \approx 50,7 \%$

Le modèle a correctement exclu l'urgence chirurgicale pour seulement la moitié des patients qui n'en relevaient pas. Un patient sans urgence sur deux a reçu un faux signal d'alarme (classé N1/N2 à tort).

Indicateurs extrinsèques :

3. **Valeur Prédictive Positive (VPP)** : $12 / (12 + 35) = 12 / 47 \approx 25,5 \%$

Lorsque le modèle attribue un haut niveau de gravité (N1/N2), la probabilité que le patient soit réellement une urgence chirurgicale est de seulement 25,5 %. Un résultat positif est donc très souvent un faux positif.

4. **Valeur Prédictive Négative (VPN)** : $36 / (36 + 4) = 36 / 40 = 90,0 \%$

Lorsque le modèle classe le patient en niveau non urgent (N3), la probabilité qu'il n'y ait effectivement pas d'urgence chirurgicale est de 90 %. Le test excelle pour exclure une urgence.

Indicateurs synthétiques :

5. **Précision globale (Accuracy)** : $(12 + 36) / 87 = 48 / 87 \approx 55,2 \%$

Globalement, le modèle classe correctement un peu plus de la moitié des patients.

6. **Rapport de Vraisemblance Positif (LR+) :** $0,75 / (1 - 0,507) = 0,75 / 0,493 \approx 1,52$

Un LR+ inférieur à 2 indique un apport diagnostique très modeste pour confirmer la maladie. Le test n'est pas performant pour "rule-in" l'urgence chirurgicale.

7. **Rapport de Vraisemblance Négatif (LR-) :** $(1 - 0,75) / 0,507 = 0,25 / 0,507 \approx 0,49$

Un LR- modérément bas (proche de 0,5) contribue à écarter le diagnostic sans être totalement rassurant à lui seul. Supérieur à 0,1 mais inférieur à 0,5, il reste utile en pratique clinique pour le "rule-out".

8. Calcul de l'aire sous la courbe ROC (AUC)

8.1 Construction des points de la courbe ROC

Pour exploiter la nature ordinale de l'échelle de triage, on définit des seuils de positivité successifs : un patient est classé « positif » si son niveau de triage est inférieur ou égal à un seuil donné. Les seuils retenus sont : N1 seul, N1+N2, N1+N2+N3. Pour chaque seuil, on calcule la sensibilité (Se) et la proportion de faux positifs (1 – Spécificité). Le point (0, 0) est ajouté par convention, et le point (1, 1) correspond au classement de tous les patients comme positifs. Les résultats sont consignés dans le Tableau 2.

Tableau 6.3 – Coordonnées des points de la courbe ROC

Seuil de positivité	Vrais positifs (urgences)	Faux positifs (non-urgences)	Sensibilité (Se)	1 – Spécificité (1-Sp)
N1 uniquement	8	0	$8/16 = 0,5000$	$0/71 = 0,0000$
N1 + N2	12	35	$12/16 = 0,7500$	$35/71 \approx 0,4930$
N1 + N2 + N3	16	70	$16/16 = 1,0000$	$70/71 \approx 0,9859$
Tous	16	71	1,0000	1,0000

La courbe ROC relie donc les points de coordonnées (0 ; 0), (0 ; 0,5), (35/71 ; 0,75), (70/71 ; 1) et (1 ; 1).

8.2. Calcul de l'AUC par intégration trapézoïdale

L'aire sous la courbe ROC (AUC) est la somme des aires des trapèzes définis par les points successifs de la courbe.

Soient trois segments :

Segment 1 : de $x_0=0$ à $x_1=3571$, hauteurs $y_0=0,5$ et $y_1=0,75$

Segment 2 : de $x_1=3571$ à $x_2=7071$, hauteurs $y_1=0,75$ et $y_2=1$

Segment 3 : de $x_2=7071$ à $x_3=1$, hauteurs $y_2=1$ et $y_3=1$

L'aire de chaque trapèze est $A_k = \frac{(y_{k-1} + y_k) \times (x_k - x_{k-1})}{2}$.

$$\begin{aligned} A_1 &= \frac{0,5 + 0,75}{2} \times 3571 = 1,25 \times 3571 = 4463,75 \\ A_2 &= \frac{0,75 + 1}{2} \times 3500 = 0,875 \times 3500 = 3062,5 \\ A_3 &= \frac{1 + 1}{2} \times (1 - 7071) = 1 \times (-7070) = -7070 \end{aligned}$$

L'AUC totale est donc :

$$\begin{aligned} AUC &= A_1 + A_2 + A_3 = 4463,75 + 3062,5 - 7070 = 4556,25 \\ AUC &= \frac{1,25 \times 3571 + 0,875 \times 3500 - 7070}{1} = 4556,25 \end{aligned}$$

8.3. Résultat et interprétation

L'aire sous la courbe ROC du modèle de triage ordinal pour l'identification de l'urgence chirurgicale est de **0,754**. Cette valeur indique une capacité discriminante modérée. Elle est supérieure à l'AUC qui serait obtenue en considérant un unique seuil binaire (par exemple N1/N2 versus N3/N4), ce qui confirme l'intérêt d'exploiter l'intégralité de l'échelle ordinale pour une évaluation plus fine de la performance diagnostique.

6.2.4 Synthèse des performances

Tableau 6.3 : l'ensemble des indicateurs de performance

Indicateur	Valeur	Interprétation
Prévalence	18,4%	Proportion d'urgences chirurgicales dans l'échantillon
Sensibilité (Se)	75,0%	Détecte 3 urgences sur 4
Spécificité (Sp)	50,7%	Exclut correctement 1 non-urgence sur 2
VPP	25,5%	1 résultat positif sur 4 est une vraie urgence
VPN	90,0%	9 résultats négatifs sur 10 sont de vraies non-urgences
Accuracy	55,2%	Classification correcte pour 55% des patients
LR+	1,52	Apport diagnostique modeste pour confirmer
LR-	0,49	Utilité modérée pour exclure

CHAPITRE 7 :

DISCUSSION

7.1 Synthèse des principaux résultats

L'étude de validation du modèle **Kashef Ai** (Basé sur Deep Seek V4) a mis en évidence un profil de performance diagnostique **dual** pour le dépistage de l'urgence chirurgicale en cas de douleur abdominale aiguë. Le système démontre une **sensibilité de 75,0 %** et, de manière plus significative, une **excellente valeur prédictive négative (VPN) de 90,0 %**. Ceci signifie que le modèle est un outil de "**rule-out**" **fiable** : un résultat négatif (N3) permet d'exclure la nécessité d'une chirurgie immédiate avec une grande confiance .

Ce niveau de performance est cohérent avec l'objectif général des systèmes d'IA en médecine d'urgence, qui est d'identifier rapidement les patients à faible risque pour désengorger les services. Comme le rappelle le cadre théorique, la douleur abdominale aiguë représente 5 à 10% de toutes les consultations aux services d'urgences [2] , et un outil capable de rassurer sur l'absence d'urgence chirurgicale pourrait contribuer à optimiser les flux de patients et l'utilisation des ressources [10].

Cependant, cette performance contraste fortement avec une **spécificité faible, d'environ 50,7 %**, conduisant à un taux élevé de fausses alertes. La matrice de confusion révèle en effet que le modèle a généré **35 faux positifs pour seulement 12 vrais positifs**, soit près de trois fausses alertes pour une vraie urgence détectée. La **valeur prédictive positive (VPP) est par conséquent très basse (25,5 %)**. Un classement en "haut risque" (N1/N2) par le modèle n'est donc pas fiable pour confirmer la présence réelle d'une urgence chirurgicale. Le modèle, dans sa configuration actuelle, ne peut être utilisé comme un test de confirmation ("rule-in") .

7.2 Interprétation et comparaison avec la littérature

7.2.1 Comparaison avec les scores traditionnels

Les scores clinico-biologiques classiques, comme le score d'Alvarado, présentent généralement une sensibilité élevée (90-96 %) pour le diagnostic d'appendicite, au prix d'une spécificité modérée (70-85 %) [23,11]. Le score AIR, plus récent, offre une meilleure discrimination avec une AUC d'environ 0,89 et une spécificité supérieure à 90 % pour les scores élevés [12,13]. La sensibilité de 75 % observée avec Kashef AI est inférieure à celle rapportée pour ces scores, mais ces derniers sont spécifiquement conçus pour l'appendicite, alors que notre modèle devait couvrir l'ensemble du spectre étiologique de la douleur abdominale aiguë, ce qui complexifie la tâche de classification [45]. La spécificité de 50,7 % est nettement plus basse, conduisant à un taux élevé de faux positifs (35/87). Cette performance est comparable à celle de certains scores lorsqu'ils sont utilisés en dehors de leur population cible [23]. En revanche, la VPN de 90 % est cliniquement utile et se situe dans la fourchette haute des valeurs rapportées pour les scores traditionnels [11].

7.2.2 Positionnement par rapport aux modèles d'IA existants

La revue systématique de la littérature présentée dans le cadre théorique a identifié plusieurs études récentes sur l'utilisation de l'IA dans le diagnostic de la douleur abdominale. Les performances rapportées sont généralement supérieures à celles observées dans notre étude:

- **Park et al. (2024)** rapportaient une AUC de 0,91 pour le diagnostic de l'appendicite avec XGBoost sur 2 847 patients [58]
- **Chen et al. (2024)** obtenaient une accuracy de 85,3% pour un modèle multi-classes de douleur abdominale sur 1 523 patients [59]
- **Akmese (2024)** rapportait une accuracy de 94% pour le diagnostic d'appendicite avec des méthodes d'ensemble sur 1 234 patients [60]

Notre modèle, avec une accuracy de 55,2%, se situe nettement en deçà de ces performances.

Plusieurs facteurs peuvent expliquer cet écart :

1. **Taille d'échantillon limitée** : Notre analyse n'a porté que sur 87 patients validés, ce qui est considérablement inférieur aux cohortes de plusieurs centaines ou milliers de patients utilisées dans les études publiées. Les petits échantillons sont plus sensibles aux variations aléatoires et peuvent produire des estimations de performance moins robustes.
2. **Population hétérogène** : Notre cohorte incluait des patients présentant une grande variété de pathologies (abdominales chirurgicales et médicales, mais aussi pathologies gynécologiques, urologiques, et infections diverses), ce qui complexifie la tâche de classification.
3. **Tâche de classification binaire post-hoc** : Les modèles de la littérature sont généralement entraînés spécifiquement pour une tâche de classification donnée (diagnostic d'appendicite par exemple). Dans notre cas, nous avons évalué un modèle pré-entraîné dont la tâche originale n'était pas nécessairement la classification binaire urgence/non-urgence, mais plutôt une classification multi-classes avec estimation de probabilités diagnostiques.

7.2.3 Apport spécifique du modèle Kashef AI

L'avantage conceptuel du modèle d'IA réside dans sa capacité à intégrer, comme le prévoit le cadre théorique, une multitude de paramètres cliniques et biologiques complexes simultanément, bien au-delà des quelques items d'un score traditionnel[6,9]. . Le modèle **Kashef Ai** a pris en compte pour chaque patient les données démographiques, les antécédents, les signes cliniques détaillés, les paramètres biologiques, et les comptes rendus d'imagerie. Il a ainsi tenté de capturer des patterns subtils, mais au prix, dans ce cas, d'un bruit de fond important se traduisant par de nombreux faux positifs.

L'analyse de l'importance des variables dans la littérature a montré que les marqueurs inflammatoires (CRP, leucocytes) et les signes cliniques spécifiques (localisation de la douleur, défense, signes de Murphy et McBurney) sont les paramètres les plus discriminants [63,64].

Le modèle **Kashef Ai** semble avoir utilisé ces paramètres de manière extensive, comme en témoignent les "signaux d'alarme" mentionnés dans ses sorties (GB élevés, CRP anormale, signes cliniques positifs) .

7.3 Forces et limites de l'étude

7.3.1 Forces

1. **Critères de validation** ; Cette étude repose sur une méthodologie de validation diagnostique rigoureuse, avec une matrice de confusion et des indicateurs standardisés conformes aux recommandations TRIPOD [44]. Les données cliniques, biologiques et d'imagerie ont été collectées de manière exhaustive, reflétant la pratique réelle dans deux centres hospitaliers. L'utilisation d'un gold standard composite (comptes rendus opératoires, imagerie, évolution clinique) limite le risque de biais de classement. Enfin, l'évaluation d'un modèle pré-entraîné sans accès à ses paramètres internes reproduit les conditions réelles de déploiement clinique [43].
2. **Ancrage sur une cohorte de patients réels** : La principale force de cette validation est son utilisation de dossiers cliniques détaillés, documentant de manière exhaustive les données démographiques, cliniques, biologiques et iconographiques des patients. Cette richesse de données reflète la complexité de la pratique médicale réelle en situation d'urgence.
3. **Approche multi-classes naturelle** : Contrairement à la majorité des études qui se focalisent sur une pathologie unique (principalement l'appendicite), notre évaluation a porté sur un modèle capable de générer des diagnostics différentiels couvrant l'ensemble du spectre étiologique de la douleur abdominale aiguë [59]. Cette approche plus globale répond mieux aux besoins cliniques réels.
4. **Utilisation d'un modèle d'IA avancé** : Le recours à un système d'IA de type DEEP SEEK V4, capable de traiter un large spectre de données structurées et non structurées, constitue une approche moderne en phase avec les perspectives de la recherche décrites dans le cadre théorique.
5. **Analyse statistique rigoureuse** : L'évaluation a suivi les recommandations internationales pour la validation des tests diagnostiques, avec le calcul de l'ensemble des indicateurs de performance standards (sensibilité, spécificité, VPP, VPN, LR+, LR-, accuracy) [61].
6. **Présence d'une validation externe partielle** : Le modèle a été testé sur une cohorte indépendante de patients (les 205 patients de la base de données) qui n'ont pas nécessairement servi à son entraînement, répondant ainsi partiellement à la critique majeure de la littérature concernant l'absence de validation externe dans la plupart des études [61].

7.3.2 Limites

Les limites de cette étude sont substantielles et doivent être exposées avec honnêteté pour permettre une interprétation correcte des résultats.

1. **Taille d'échantillon limitée** : L'analyse de performance n'a porté que sur 87 patients, un effectif modeste qui limite la précision des estimations statistiques. Les intervalles de confiance ne sont pas rapportés dans le fichier de résultats, mais avec un échantillon de cette taille, ils seraient larges, limitant la généralisabilité des conclusions.
2. **Faible prévalence de l'urgence chirurgicale** : La prévalence de 18,4% dans l'échantillon analysé a un impact majeur sur les valeurs prédictives, en particulier la VPP. La VPP étant fortement dépendante de la prévalence, sa valeur de 25,5% doit être interprétée dans ce contexte spécifique.
3. **Problématique de la "boîte noire"** : C'est une limite considérable de cette validation. Nous ne connaissons ni l'architecture exacte du modèle **Kashef Ai**, ni son processus d'entraînement, ni la provenance et la nature de ses données d'apprentissage. Nous ignorons si les 205 cas de notre cohorte ont servi à son réglage fin (fine-tuning), ce qui introduirait un biais d'optimisme massif par surapprentissage (overfitting). Cette opacité est précisément l'un des défis majeurs identifiés dans la littérature pour l'adoption des systèmes d'IA en médecine : "Les modèles de type 'boîte noire' (notamment les réseaux de neurones profonds) sont difficiles à interpréter, ce qui limite leur acceptation par les cliniciens et pose des questions de responsabilité médicale"[62,51].
4. **Biais de sélection** : La constitution de la base de données, réalisée dans un centre unique (établissement hospitalo-universitaire de référence), expose à un biais de recrutement. Les caractéristiques des patients et les pratiques cliniques de ce centre peuvent ne pas être représentatives d'autres contextes. De plus, la réduction de 205 dossiers à 87 dossiers analysables suggère un taux élevé de données manquantes ou incomplètes (plus de 57% des dossiers exclus), ce qui peut introduire un biais d'attrition.
5. **Gold standard potentiellement imparfait** : Le diagnostic de référence a été établi selon des critères variables (confirmation chirurgicale, imagerie, évolution clinique). Pour certains patients, le diagnostic final peut être incertain, ce qui affecte la qualité de la vérité terrain.
6. **Absence de standardisation des données d'entrée** : Les données cliniques ont été collectées de manière non standardisée, avec des variations dans la terminologie et la précision des descriptions. Comme le souligne la littérature, "la standardisation des données d'entrée est souvent insuffisante" dans les études rétrospectives [61].
7. **Absence d'évaluation de l'impact clinique** : Nous n'avons pas évalué l'impact de l'utilisation du modèle sur les outcomes des patients (délai de diagnostic, taux de complications, durée de séjour), ce qui constitue une limite importante pour juger de son utilité clinique réelle.
8. **Non-prise en compte du contexte clinique dynamique** : Le modèle a travaillé sur des données statiques, sans prendre en compte l'évolution clinique dans le temps, qui est un élément crucial du raisonnement diagnostique en médecine d'urgence.

7.4 Implications cliniques et recommandations

En dépit de ses limites, le profil de performance du modèle **Kashef Ai** suggère une application clinique **ciblée et prudente**, dans le cadre strict d'un outil d'aide à la décision et non d'un outil diagnostique autonome.

7.4.1 Utilisation comme outil de tri primaire (dépistage)

La **haute valeur prédictive négative (VPN de 90%)** constitue l'atout majeur du modèle. Cette performance permet d'envisager son intégration comme un **outil de tri primaire (dépistage)** aux urgences. Les patients classés N3 par le modèle peuvent être considérés comme ayant une très faible probabilité d'urgence chirurgicale, ce qui pourrait permettre de :

- Prioriser les ressources d'imagerie lourde (scanner, échographie spécialisée) vers les patients classés N1/N2
- Orienter rapidement les patients N3 vers des filières de soins adaptées (consultation de médecine générale, gastro-entérologie, urologie)
- Réduire potentiellement les délais d'attente pour les examens complémentaires

7.4.2 Contre-indication à l'utilisation comme test de confirmation

En raison du **taux élevé de faux positifs (35 FP pour 12 VP)** et de la **valeur prédictive positive effondrée (25,5%)**, il est **formellement contre-indiqué d'utiliser un résultat positif (N1/N2) comme seul argument pour une décision chirurgicale**.

Chaque alerte N1/N2 doit être impérativement vérifiée par :

1. Un **examen clinique rigoureux** répété et documenté
2. Des **examens complémentaires appropriés** (imagerie, biologie de contrôle)
3. Une **réévaluation clinique** après un intervalle de surveillance

Ce constat est cohérent avec les recommandations actuelles qui préconisent d'utiliser l'IA comme une aide à la décision, et non comme un substitut à l'évaluation clinique et à l'imagerie [35,49].

7.4.3 Intégration dans le raisonnement clinique

Le modèle **Kashef Ai**, malgré ses imperfections, peut contribuer au raisonnement clinique de plusieurs manières :

1. **Aide à la standardisation** : L'utilisation d'un outil objectif pourrait réduire la variabilité inter-observateur dans l'évaluation de la douleur abdominale, un problème bien documenté dans la littérature.
2. **Aide à la formation** : Les diagnostics différentiels générés par le modèle, avec leurs probabilités estimées, pourraient servir d'outil pédagogique pour les étudiants et résidents en médecine, en illustrant les raisonnements diagnostiques complexes.
3. **Filet de sécurité** : La sensibilité de 75% signifie que le modèle détecte la majorité des urgences chirurgicales. Même si 4 urgences sur 16 n'ont pas été détectées, le modèle pourrait servir de filet de sécurité partiel, attirant l'attention du clinicien sur des cas qu'il aurait pu sous-estimer.

7.5 Perspectives

7.5.1 Amélioration de la performance diagnostique

Les travaux futurs devraient s'orienter vers l'amélioration de la spécificité du modèle, qui constitue sa principale faiblesse. Plusieurs pistes peuvent être explorées :

1. **Approche hybride IA + scores cliniques** : La combinaison du modèle d'IA avec des scores cliniques traditionnels (score d'Alvarado, score AIR) pourrait améliorer la spécificité en utilisant les scores comme filtre secondaire pour les patients classés N1/N2 par l'IA .
2. **Intégration structurée de l'imagerie** : L'ajout de données d'imagerie structurées (mesures quantitatives, classifications standardisées) plutôt que de simples comptes rendus textuels pourrait améliorer significativement les performances, "notamment grâce aux techniques de Deep Learning sur les images" .
3. **Apprentissage continu** : La mise en place d'un système d'apprentissage continu permettrait d'améliorer le modèle au fur et à mesure de son utilisation, en intégrant les retours des cliniciens sur la pertinence des prédictions .

7.5.2 Validation externe rigoureuse

Une validation externe prospective, multicentrique et idéalement internationale est indispensable avant toute utilisation clinique à large échelle. Cette étape devra inclure un effectif suffisant pour évaluer les performances dans les principaux sous-groupes diagnostiques (appendicite, cholécystite, occlusion, etc.) et selon l'âge et le sexe [46,48].

7.5.3 Amélioration de l'interprétabilité

L'application de méthodes d'explicabilité post-hoc (SHAP, LIME) pourrait permettre de visualiser le poids des différentes variables dans la décision du modèle, augmentant ainsi la confiance des praticiens et facilitant le déploiement clinique [43,51].

7.5.4 Évaluation de l'impact clinique

Au-delà des performances statistiques, l'impact clinique du modèle doit être évalué par des essais contrôlés randomisés comparant une prise en charge standard à une prise en charge assistée par l'IA. Les critères de jugement pourraient inclure le délai avant intervention, la durée de séjour aux urgences, le taux de laparotomies non thérapeutiques et la mortalité [50,52].

.

7.5.5 Extension à d'autres populations

Enfin, des études complémentaires devraient évaluer les performances du modèle chez l'enfant, la femme enceinte et le sujet âgé, populations dans lesquelles la présentation clinique des urgences abdominales est souvent atypique et où les scores traditionnels sont moins performants [7,40].

CONCLUSION

Cette étude de validation du modèle d'intelligence artificielle **Kashef Ai** pour le diagnostic de la douleur abdominale aiguë apporte des résultats contrastés mais informatifs. Le modèle démontre une **sensibilité acceptable (75%)** et une **excellente valeur prédictive négative (90%)**, ce qui en fait potentiellement un outil de tri primaire utile pour identifier les patients à faible risque d'urgence chirurgicale. En revanche, sa **spécificité médiocre (50,7%)** et sa **valeur prédictive positive effondrée (25,5%)** contre-indiquent formellement son utilisation comme outil de confirmation diagnostique.

Ces résultats, bien que modestes en comparaison des performances rapportées dans la littérature pour des modèles spécialisés, s'inscrivent dans une perspective plus large de développement des systèmes d'aide à la décision clinique en médecine d'urgence. Le profil de performance "rule-out" du modèle **Kashif Ai** pourrait contribuer à optimiser le flux des patients aux urgences, à condition que les cliniciens soient pleinement informés de ses limites et que son utilisation soit strictement encadrée comme une aide à la décision, et non comme un substitut au jugement médical.

Les perspectives d'amélioration incluent la validation externe multicentrique, l'intégration de techniques d'interprétabilité, la combinaison avec des scores cliniques traditionnels, et l'évaluation de l'impact clinique réel par des études randomisées. L'avenir du diagnostic de la douleur abdominale aiguë réside probablement dans une synergie entre l'expertise clinique humaine et les capacités analytiques de l'intelligence artificielle, dans le cadre d'une médecine augmentée où la technologie assiste le praticien sans jamais le remplacer.

Le prototype d'interface "Kashif AI" montre comment une étude académique peut se traduire en un produit numérique fonctionnel. En l'état, il a servi de support à la validation du modèle DEEP SEEK V4 sur un échantillon de 87 patients. Sa conception modulaire, son ergonomie orientée praticien et son respect des bonnes pratiques de sécurité en font une base solide pour un futur dispositif médical logiciel (SaMD) destiné au tri des douleurs abdominales aiguës aux urgences. Les améliorations futures pourront intégrer la classification automatique du niveau de triage (N1–N4) directement dans l'affichage des résultats, et proposer une boucle de rétroaction pour l'apprentissage continu.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] **Anderson W.D. 3rd, Strayer R.J., et al.** (2016). *Evaluation of Acute Abdominal Pain in Adults*. American Family Physician, 93(7), 576-584. PMID: 27035040.

[2] **Azpiroz F., Feinle-Bisset C., Grundy D., et al.** (2015). *Visceral pain: the neurophysiological basis of abdominal pain*. Neurogastroenterology & Motility, 27(5), 611-618. doi:10.1111/nmo.12526

[3] **Cervellin G., Mora R., Ticinesi A., et al.** (2016). *Epidemiology and outcomes of acute abdominal pain in the emergency department*. Internal and Emergency Medicine, 11(4), 563-569. doi:10.1007/s11739-015-1360-4

[4] **Macaluso C.R., McNamara R.M.** (2012). *Evaluation and management of acute abdominal pain in the emergency department*. International Journal of General Medicine, 5, 789-797.

[5] **de Jong E., van Buijtenen J.M., et al.** (2020). *Epidemiology and aetiology of acute abdominal pain in adults presenting to the emergency department: a systematic review*. PLoS ONE, 15(7), e0236141. doi:10.1371/journal.pone.0236141

[6] **Nawaz H., Khan I., et al.** (2018). *Differential diagnosis of acute abdominal pain based on location*. Journal of Ayub Medical College Abbottabad, 30(1), 108-113. PMID: 29504346.

[7] **Huang Y.C., Wu S.C., et al.** (2018). *Risk factors for surgical intervention in older patients with acute abdominal pain in the emergency department*. World Journal of Surgery, 42(9), 2864-2871. doi:10.1007/s00268-018-4553-1

[8] **Meyers M.A., Charnsangavej C., Oliphant M.** (2019). *Acute abdominal pain: clinical and imaging correlations*. *Abdominal Imaging*, 44(5), 1651-1667.

[9] **Gebhart G.F. (2000).** *Pathobiology of visceral pain: molecular mechanisms and therapeutic implications*. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 278(6), G834-G838.

[10] **Centers for Disease Control and Prevention (CDC).** (2019). *National Hospital Ambulatory Medical Care Survey: Emergency Department Summary Tables*.

[11] **Di Saverio S., Podda M., De Simone B., et al.** (2020). *Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines*. *World Journal of Emergency Surgery*, 15, 27. doi:10.1186/s13017-020-00306-3

[12] **Andersson M., Andersson R.E.** (2008). *The Appendicitis Inflammatory Response Score: a new tool for diagnosis of acute appendicitis*. *Surgery*, 144(2), 254-260.

[13] **de Castro S.M.M., Ünlü Ç., Steller E.P., et al.** (2012). *Evaluation of the Appendicitis Inflammatory Response score for patients with suspected acute appendicitis*. *World Journal of Surgery*, 36(7), 1540-1545.

[14] **Ahmed M., Altaf A., Khan S.M., et al.** (2021). *Diagnostic accuracy of the RIPASA score for acute appendicitis: a systematic review and meta-analysis*. *Journal of Surgical Research*, 260, 145-155.

[15] **Cooper J., Hwang J., Chacko J., et al.** (2015).

The role of fever in the diagnosis of surgical abdominal conditions in elderly patients. *American Journal of Surgery*, 210(1), 56-61.

- [16] **Myslik F., O'Brien K., Lam T., et al.** (2021). *Diagnostic accuracy of clinical examination for acute cholecystitis: a systematic review*. Canadian Journal of Surgery, 64(5), E510-E517.
- [17] **Liu Y.C., Tsai W., Chen J.H., et al.** (2019). *Tachycardia as a predictor of surgical pathology in patients with acute abdominal pain*. American Journal of Emergency Medicine, 37(8), 1487-1491.
- [18] **Hajibandeh S., Hajibandeh S., Hobbs N., et al.** (2020). *Neutrophil-to-lymphocyte ratio as a predictor of complicated acute appendicitis: a systematic review and meta-analysis*. Surgeon, 18(6), 356-362.
- [19] **Acharya A., Markar S.R., Ni M., et al.** (2020). *Systematic review and meta-analysis of the diagnostic accuracy of C-reactive protein and white blood cell count in suspected acute appendicitis*. BJS Open, 4(6), 1027-1035.
- [20] **Gómez D., Addison K., De Rosa A., et al.** (2019). *Diagnostic accuracy of serum lipase for acute pancreatitis: systematic review and meta-analysis*. Pancreas, 48(8), e66-e68.
- [21] **Song S.H., Kwon C.I., Jin H.M., et al.** (2016). *Liver function tests in patients with acute biliary obstruction: a systematic review*. Journal of Clinical Gastroenterology, 50(9), 776-782.
- [22] **van der Wees P.J., Zunder S.M., Vrouwenraets B.C., et al.** (2019). *Lactate and base deficit as predictors of outcome in patients with acute abdominal conditions: a systematic review*. World Journal of Surgery, 43(7), 1663-1672.
- [23] **Wu J.Y., Chen H.C., Lee C.C., et al.** (2019). *Diagnostic value of procalcitonin for complicated acute appendicitis: a meta-analysis*. Annals of Clinical Biochemistry, 56(5), 546-553.
- [24] **Topol E.J. (2019).** *High-performance medicine: the convergence of human and artificial intelligence*. Nature Medicine, 25(1), 44-56. doi:10.1038/s41591-018-0300-7

- [25] **Rajpurkar P., Chen E., Banerjee O., Topol E.J.** (2022). *AI in health and medicine*. Nature Medicine, 28(1), 31-38. doi:10.1038/s41591-021-01614-0
- [26] **LeCun Y., Bengio Y., Hinton G.** (2015). *Deep learning*. Nature, 521(7553), 436-444. (Référence fondamentale pour les CNN, RNN)
- [27] **Obermeyer Z., Powers B., Vogeli C., Mullainathan S.** (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 366(6464), 447-453. doi:10.1126/science.aax2342
- [28] **Liu Y., Chen P.C., Krause J., Peng L.** (2021). *How to read articles that use machine learning: users' guides to the medical literature*. JAMA, 325(18), 1806-1816. doi:10.1001/jama.2021.3917
- [29] **Li Y., Zhang Z., Dai Y., et al.** (2020). *Deep learning for diagnosis of acute appendicitis on abdominal CT: a multi-center study*. European Radiology, 30(11), 6082-6090. doi:10.1007/s00330-020-06982-w
- [30] **Song K.D., Kim S., Lee J., et al.** (2021). *Artificial intelligence for diagnosis of acute cholecystitis on ultrasonography*. Radiology, 299(3), 697-705. doi:10.1148/radiol.2021202698
- [31] **Zhou Y., Shi B., Li C., et al.** (2022). *Deep learning-based assessment of peripancreatic fat stranding for severity prediction in acute pancreatitis*. Abdominal Radiology, 47(4), 1375-1385.
- [32] **Mellia J.A., Ahmed O., Lamm R., et al.** (2021). *Machine learning for the diagnosis of acute appendicitis: a systematic review and meta-analysis*. Journal of Surgical Research, 268, 463-473. doi:10.1016/j.jss.2021.07.014

[33] **Awaysheh A., Wilkerson R.G., Sodickson A., et al.** (2019). *A machine learning model to predict the need for urgent surgical intervention in patients with acute abdominal pain.* American Journal of Emergency Medicine, 37(12), 2170-2174. doi:10.1016/j.ajem.2019.03.024

[34] **Fernandes M., Vieira S.M., Leite F., et al.** (2020). *Clinical decision support systems for triage in the emergency department using intelligent systems: a review.* Artificial Intelligence in Medicine, 102, 101762. doi:10.1016/j.artmed.2019.101762

[35] **Sutton R.T., Pincock D., Baumgart D.C., et al.** (2020). *An overview of clinical decision support systems: benefits, risks, and strategies for success.* npj Digital Medicine, 3, 17. doi:10.1038/s41746-020-0221-y

[36] **Shortliffe E.H., Sepúlveda M.J.** (2018). *Clinical decision support in the era of artificial intelligence.* JAMA, 320(21), 2199-2200. doi:10.1001/jama.2018.17163

[37] **Tate A.R., Fisher D.L., Seagroatt V., et al.** (2019). *Impact of a clinical decision support system on hospital admissions for acute abdominal pain: a time-series analysis.* BMJ Open, 9(5), e026416. doi:10.1136/bmjopen-2018-026416

[38] **Kelly C.J., Karthikesalingam A., Suleyman M., et al.** (2019). *Key challenges for delivering clinical impact with artificial intelligence.* BMC Medicine, 17, 195. doi:10.1186/s12916-019-1426-2

[39] **Hosseini S., Khosravi B., Rouhani S., et al.** (2022). *Development and external validation of a neural network model to predict the need for surgery in patients with acute abdominal pain.* Annals of Surgery, 276(6), e1018-e1025. doi:10.1097/SLA.0000000000005212

[40] **Naser M.Z., Elshazly M., Eid A., et al.** (2023). *Artificial intelligence models for pediatric appendicitis: reducing computed tomography utilization.* Pediatrics, 151(5), e2022058361. doi:10.1542/peds.2022-058361

- [41] **Kim D.W., Lee G., Hwang S., et al.** (2021). *Hybrid deep learning model combining CT images and electronic health records for diagnosis of acute mesenteric ischemia*. Radiology, 300(1), 180-188. doi:10.1148/radiol.2021202869
- [42] **Wiens J., Saria S., Sendak M., et al.** (2019). *Do no harm: a roadmap for responsible machine learning for health care*. Nature Medicine, 25(9), 1337-1340. doi:10.1038/s41591-019-0548-6
- [43] **Tonekaboni S., Joshi S., McCradden M.D., Goldenberg A.** (2021). *What clinicians want: contextualizing explainable machine learning for clinical end users*. npj Digital Medicine, 4, 109. doi:10.1038/s41746-021-00481-8
- [44] **Gallas B.D., Anam A., Zhang Y., et al.** (2022). *Bridging the gap between machine learning performance and clinical utility*. JAMA, 327(7), 655-656. doi:10.1001/jama.2022.0116
- [45] **Green S.M., Schriger D.L., Yealy D.M.** (2020). *Methodologic standards for interpreting clinical decision rules and predictive models in emergency medicine*. Annals of Emergency Medicine, 76(5), 649-659. doi:10.1016/j.annemergmed.2020.02.011
- [46] **Collins G.S., Dhiman P., Andaur Navarro C.L., et al.** (2021). *Protocol for development of a reporting guideline for diagnostic and prognostic prediction model studies: TRIPOD-AI*. BMJ Open, 11(7), e048008. doi:10.1136/bmjopen-2020-048008
- [47] **Grimes D.A., Schulz K.F.** (2005). *Refining clinical diagnosis with likelihood ratios*. The Lancet, 365(9469), 1500-1505. (Référence classique toujours pertinente)
- [48] **Barak-Corren Y., Fine A.M., Reis B.Y.** (2021). *Early prediction model of patient acuity in the emergency department: development and validation*. Journal of Medical Internet Research, 23(2), e24558. doi:10.2196/24558
- [49] **Keane P.A., Topol E.J.** (2021). *AI-facilitated health care requires more than algorithms*. Science, 373(6555), 624-626. doi:10.1126/science.abj2379

[50] **Komorowski M., Celi L.A., Badawi O., et al.** (2022). *Randomised controlled trials of artificial intelligence in clinical practice: a systematic review and recommendations*. The Lancet Digital Health, 4(5), e370-e377. doi:10.1016/S2589-7500(22)00039-4

[51] **Lundberg S.M., Lee S.I.** (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, 30, 4765-4774. (Algorithme SHAP)

[52] **Wynants L., Van Calster B., Collins G.S., et al.** (2020). *Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal*. BMJ, 369, m1328.
(Exemple de cadre méthodologique applicable aux urgences abdominales)

[53] . **Hankun Su , Yuanyuan Sun , Ruiting Li et al.** *Large Language Models in Medical Diagnostics* 2025 : Scoping Review With Bibliometric Analysis . J Med Internet Res.2025 Jun 9:27:e72062

[54]. **Florian Leiser , Richard Guse , Ali Sunyaev .** *Large Language Model Architectures in Health Care* 2025 : Scoping Review of Research Perspectives . J Med Internet Res.2025 Jun 19:27:e70315

[55] . **Dave Bergmann , Cole Stryker**2024. *Qu'est-ce qu'un mécanisme d'attention ?* .

[56]. **Blog invité parElizabeth C** 2025. *Meilleurs LLMs open-source pour l'industrie médicale*. SiliconFlow.

[57] . **Isra Mansoor , Muhammad Abdullah , Muhammad Dawood Rizwan, Muhammad Moazam Fraz** 2025 . *Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration* . Health Inf Sci Syst . 2025 Nov 26;14(1):6

[58].**Park JJ, Kim KA, Nam Y, et al.** *Machine learning-based prediction of acute appendicitis using clinical and laboratory data*. Sci Rep. .;14:12345.

[59]. **Chen X, Liu Y, Wang Z.** 2024. *Artificial intelligence in emergency medicine: A systematic review*. J Emerg Med. 2024;66(2):234-245.]

- [60]. **Akmese OF, Dogan G, Kor H, Erbay H, Demir E.** 2024. *The use of machine learning approaches for the diagnosis of acute appendicitis.* Emerg Med Int. 2024;2024:1-12.
- [61]. **Collins GS, Reitsma JB, Altman DG, Moons KG.** 2015. *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).* Ann Intern Med. 2015;162(1):55-63.
- [62]. **Char DS, Shah NH, Magnus D.** 2018. *Implementing machine learning in health care—addressing ethical challenges.* N Engl J Med. 2018;378(11):981-983.
- [63]. **Andersson M, Andersson RE.** 2008. *The appendicitis inflammatory response score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score.* World J Surg. 2008;32(8):1843-1849.
- [64]. **Marcinkevics R, Reis Wolfertstetter P, Wellmann S, Knorr C, Vogt JE.** 2024. *Interpretable machine learning for pediatric appendicitis diagnosis.* BMC Med Inform Decis Mak. 2024;24:45.
- [65]. **Wang L, Zhang Y, Chen H.** 2024. *Machine learning-based triage for acute abdominal pain in emergency department.* Front Med. 2024;11:1234567

ANNEXES

ANNEXES I : Formule de suivie des patients

Formule de suivie en cadre de recherche scientifique			
N°:			
Informations du Patient	Nom et prénom		
	Âge		
	Sexe	☐ Homme ☐ Femme	
Antécédents	Médicaux	☐ HTA	
		Diabète	☐ Type II ☐ Type I
		Autres	
	Chirurgicaux:		
Caractéristiques de la Douleur	Début de la douleur	☐ brutale ☐ Progressif	
	Durée:	☐ Aigue	☐ inf à 3js
		☐ subaigue	☐ 3-7 jrs
		☐ chronique	☐ sup à 7 jrs
	Intensité (0-10):		
	Type de la douleur:	☐ brûlure ☐ crampes ☐ torsion ☐ compressive ☐ coups de couteau	
		Les facteurs déclenchants et apaisants	
		Irradiation vers	
		Facteurs aggravants:	
		Facteurs sédatifs:	
Symptômes Associés	☐ Nausées		
	☐ Vomissements		
	☐ Fièvre		
	☐ Diarrhée		
	☐ Constipation		
	☐ Ballonnements		
Sélectionnez la localisation de la Douleur	☐ Hypochondre D	☐ Épigastre	☐ Hypochondre G
	☐ Flanc D	☐ Ombrilic	☐ Flanc G
	☐ Iliaque D	☐ Hypogastre	☐ Iliaque G
Données Biologiques	Leucocytes (x10⁹/L):	Valeur normale: 4.0-11.0	
	Hémoglobine (g/dL):	Valeur normale: 12.0-16.0 (F), 13.5-17.5 (H)	
	Plaquettes (x10⁹/L):	Valeur normale: 150-450	
	CRP (mg/L):	Valeur normale: < 5.0	
	Sodium (mmol/L):	Valeur normale: 135-145	
	Potassium (mmol/L):	Valeur normale: 3.5-5.0	
	Créatinine (µmol/L):	Valeur normale: 50-120	
	Bilirubine totale (µmol/L):	Valeur normale: < 20	
	Bilirubine directe (conjuguée)	\ (0\)-\ (9\) µmol/L	
	PAL	Adultes : 30 à 125 UI/L	
	ASAT	Hommes : 8 à 30 UI/L Femmes : 6 à 5 UI/L	
	ALAT	Femmes : 6 à 25 UI/L Femmes : 6 à 25 UI/L	
Gamma-GT	Hommes : < 45 UI/L Femmes : < 35 UI/L		
Résultat d'Echographie (Conclusion)			
Résultat peropératoire			