الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Frères Mentouri Constantine 1
Faculté des Sciences de la Nature et de la Vie

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Département de Biologie Appliquée

قسم البيولوجيا التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master
Domaine : Sciences de la Nature et de la Vie
Filière : Biotechnologie
Spécialité : Bio-informatique

THÈME :

Approche par IA supervisée dans la prédiction de l'infertilité masculine

Présenté par :

BOUBIA AMAR MOUNIB

HABIBATNI CHEMSEDDINE

MEZZAR ABDELAZIZ

Soutenu le : 25 - 06 - 2025

Devant le jury :

Président du jury : Dr. BELLIL INES (Pr.)

Examinateur : Dr. DJEZZAR NEDJMA (MCB)

Encadreur : Dr. BENSAADA Mostafa (MCA)

Année universitaire 2024/2025

REMERCIEMENTS

Avant tout, nous remercions « DIEU » le tout puissant et miséricordieux, qui nous a donné la force le courage et la patience pour accomplir ce travail.

Nos vifs remerciements vont également aux membres du jury **Dr. BELLIL INES & Dr. DJEZZAR NEDJMA.** Pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'évaluer notre mémoire de fin d'études et de l'enrichir par leur contribution.

Docteur **BENSAADA Mostafa**, notre directeur du projet de fin d'études et encadreur,

Nous avons eu le grand plaisir de travailler sous votre encadrement, on a eu le privilège de vos vastes connaissances et riche expérience, vous nous avez prêté main forte depuis le premier jour et tout au long de la réalisation et rédaction, vous avez inculqué en nous l'esprit de l'organisation, vous nous avez poussé à nous surpasser, merci pour les conseils que vous nous avez prodigués, pour le temps précieux que vous nous avez accordé et votre implication. Vous nous avez accueilli et vous avez travaillé avec nous en dehors des heures de travail, nous vous admirons pour votre enthousiasme et votre amabilité, nous espérons que notre travail soit à la hauteur de l'image prestigieuse de votre encadrement, recevez notre plus profond respect et salutations les plus distinguées. Puisse ce travail sera à la hauteur de vos attentes.

Docteur **CHEHILI HAMZA** notre deuxième directeur de mémoire Nous tenons à exprimer notre profonde gratitude au Dr **CHEHILI HAMZA**, pour son précieux soutien et son expertise en informatique qui ont été indispensables à la réalisation de ce projet. Ses conseils avisés, sa patience et sa disponibilité ont grandement facilité la résolution des défis techniques rencontrés. Merci pour son engagement et son accompagnement tout au long de ce travail.

Doctorant **BEN DAHMANE ABDELHAFEDH**, Merci à vous, vous nous avez accueilli de la meilleure des façons dès le premier jour, nous avons eu la chance de bénéficier de votre expertise de vos retours d'expérience. Merci pour votre temps votre disponibilité et de nous avoir toujours répondu favorablement car on manquait de recul face aux questions soulevées par nos travaux, nous vous admirions beaucoup pour votre grand et brillant esprit de partage scientifique. Vous avez toute notre gratitude et notre profond respect.

Nous remercions également tous nos enseignants qui ont assuré notre formation tout au long de ces années. Nous soulignons que leurs efforts nous ont été bénéfiques pour effectuer ce travail.

Un grand merci à nos amis pour leur sympathie et leurs soutiens.

Nous remercions nos parents, nos frères et sœurs, sans eux, ce travail n'aurait jamais pu être réalisé. Un énorme merci à nos familles pour leur soutien indéfectible et leurs encouragements tout au long de l'élaboration de notre travail.

DEDICACES

Je dédie ce mémoire à mes chers parents,
pour leur amour inconditionnel, leurs sacrifices silencieux et leur soutien sans faille.
Leur confiance et leurs prières ont été ma force et ma lumière dans ce parcours.

À toute ma famille, pour leur affection et leur accompagnement constant.

À mes enseignants et encadrants, pour leur bienveillance, leurs conseils précieux et leur patience.

À mes amis sincères, pour leur présence, leurs encouragements et leur soutien moral.

À tous ceux qui, de près ou de loin, ont cru en moi et ont contribué à l'accomplissement de ce travail.

Que ce modeste travail soit le reflet de ma reconnaissance et de mon profond respect.

RÉSUMÉ

Ce travail de recherche explore l'application de l'intelligence artificielle (IA) dans la prédiction de l'infertilité masculine, en utilisant comme marqueurs : les niveaux de trois hormones (FSH, LH et testostérone), le volume testiculaire, ainsi que les résultats de la biopsie en cas d'azoospermie sévère. Un modèle d'apprentissage supervisé basé sur l'algorithme Random Forest, Le modèle a été conçu pour classer les patients infertiles selon les résultats de leur biopsie testiculaire : présence ou absence de spermatozoïdes..

Les résultats montrent que la FSH, la LH et le volume testiculaire constituent des marqueurs prédictifs clés, avec des performances modérées mais statistiquement significatives. Les principales limites de notre modèle concernent la taille restreinte de l'échantillon et l'absence de certaines variables moléculaires ou génétiques.

Nos résultats ouvrent néanmoins des perspectives prometteuses, notamment par l'intégration future de nouveaux biomarqueurs moléculaires (analyses génétiques, profils transcriptomiques) et par l'utilisation d'algorithmes plus avancés, en vue d'améliorer la précision diagnostique et de limiter le recours aux explorations invasives.

Mots clés : Azoospermie ; Biomarqueurs ; Classification ; Infertilité Masculine ; Intelligence Artificielle ; Prédiction.

ABSTRACT

This study explores the application of artificial intelligence (AI) in predicting male infertility, using biomarkers such as the levels of three hormones (FSH, LH, and testosterone), testicular volume, and biopsy results in cases of severe azoospermia. A supervised learning model based on the Random Forest algorithm was employed to classify patients according to their clinical and biological characteristics.

The findings show that FSH, LH, and testicular volume are key predictive markers, with moderate but statistically significant performance. The main limitations of our model lie in the small sample size and the lack of molecular or genetic biomarkers.

Nevertheless, these results offer promising perspectives, especially through the future inclusion of genetic analysis and the use of more advanced algorithms to enhance diagnostic accuracy and reduce the need for invasive procedures.

Keywords: Azoospermia; Biomarkers; Classification; Male infertility; Artificial intelligence; Prediction.

الملخص

يستكشف هذا البحث تطبيق الذكاء الاصطناعي في التنبؤ بالعقم عند الذكور، وذلك باستخدام مؤشرات حيوية تشمل مستويات ثلاث هرمونات(FSH ، LH، والتستوستيرون(، وحجم الخصيتين، ونتائج الخزعة في حالات انعدام النطاف الشديد. تم استخدام نموذج تعلم إشرافي يعتمد على خوارزمية "الغابة العشوائية (Random Forest) "لتصنيف المرضى بناءً على خصائصهم السريرية والبيولوجية.

أظهرت النتائج أن FSH و LHوحجم الخصية تُعد من أهم المؤشرات التنبؤية، مع أداء متوسط لكنه ذو دلالة إحصائية واضحة. تكمن أبرز محدوديات النموذج في حجم العينة الصغير وغياب بعض المؤشرات الجزيئية أو الوراثية.

رغم ذلك، تفتح هذه النتائج آفاقاً واعدة، خاصةً عند إدماج تحاليل وراثية مستقبلية واستخدام خوارزميات أكثر تقدماً، بهدف تحسين دقة التشخيص وتقليل الحاجة إلى الإجراءات التداخلية

الكلمات المفتاحية: انعدام الحيوانات المنوية؛ العلامات البيولوجية؛ التصنيف؛ العقم الذكوري؛ الذكاء الاصطناعي؛ التنبؤ.

LISTES DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

- **FSH :** Follicle Stimulating Hormone (Hormone folliculo-stimulante)
- **LH :** Luteinizing Hormone (Hormone lutéinisante)
- **TESTO :** Testosterone (Testostérone)
- **TG :** Taux de Gonadotrophines
- **GnRH :** Gonadotropin-Releasing Hormone (Hormone de libération des gonadotrophines)
- **AMH :** Anti-Müllerian Hormone (Hormone anti-müllérienne)
- **ABP :** Androgen Binding Protein (Protéine de liaison aux androgènes)
- **AZF :** Azoospermia Factor (Facteur d'azoospermie)
- **DAZ :** Deleted in Azoospermia
- **CBAVD :** Congenital Bilateral Absence of the Vas Deferens (Absence bilatérale congénitale des canaux déférents)
- **CFTR :** Cystic Fibrosis Transmembrane Conductance Regulator
- **AAS :** Anticorps Anti-Spermatozoïdes
- **HHC :** Hypogonadisme Hypogonadotrope Congénital
- **SCOs** : Sertoli Cell Only (Syndrome des cellules de Sertoli seules)
- **ECHO_TD :** Échographie testiculaire
- **ARRET_DE_M :** Arrêt de la maturation
- **HYPOSITIFPERM :** Hypospermatogenèse avec perméabilité partielle
- **RESULTAT_BT :** Résultat de la biopsie testiculaire
- **AI / IA :** Artificial Intelligence / Intelligence Artificielle
- **ML :** Machine Learning (Apprentissage automatique)
- **DL :** Deep Learning (Apprentissage profond)
- **PMA :** Procréation Médicalement Assistée

- **ROC :** Receiver Operating Characteristic (Courbe ROC)
- **AUC :** Area Under the Curve (Aire sous la courbe)
- **SMOTE :** Synthetic Minority Over-sampling Technique
- **CNN :** Convolutional Neural Network (Réseau neuronal convolutif)
- **ECG :** Electrocardiogram (Électrocardiogramme)
- **IST / MST :** Infections Sexuellement Transmissibles / Maladies Sexuellement Transmissibles
- **DNA / ADN :** Deoxyribonucleic Acid / Acide Désoxyribonucléique
- **KNN :** K-Nearest Neighbors (Plus proches voisins)
- **API :** Application Programming Interface (Interface de programmation d'application)

TABLE DE MATIÈRES

TABLE DES MATIÈRES

Introduction

INTRODUCTION

L'infertilité masculine représente un enjeu de santé publique important, affectant la qualité de vie des couples et nécessitant une prise en charge médicale, psychologique et sociale adaptée. Perçue comme une fatalité, elle engendre une souffrance individuelle et peut devenir un facteur de fragilisation des unions matrimoniales (Boushaba & Belaaloui, 2015). Dans les sociétés africaines, et plus particulièrement en Algérie, où la procréation demeure l'un des fondements du mariage , l'infertilité masculine génère des répercussions morales profondes sur l'individu, la famille et la structure sociale (Belhachemi et al., 2020).

Longtemps attribuée exclusivement aux femmes, l'infertilité conjugale a été interprétée à travers un prisme culturel et biomédical biaisé, reléguant la responsabilité masculine au second plan. Ce n'est qu'au XXe siècle, avec l'essor de l'andrologie et l'introduction de techniques telles que l'analyse du sperme, que l'infertilité masculine a été médicalement reconnue comme une cause indépendante de stérilité (Agarwal et al., 2021).

Aujourd'hui, selon l'Organisation mondiale de la santé, l'infertilité est définie comme l'absence de grossesse après 12 à 24 mois de rapports sexuels réguliers non protégés (WHO, 2024). Le facteur masculin est impliqué seul ou en association dans environ 50 % des cas (Pozzi et al., 2023). Néanmoins, l'évaluation de la fertilité masculine reste complexe : des paramètres spermatiques altérés peuvent être observés chez des hommes fertiles, tandis que certains hommes ayant un spermogramme normal peuvent présenter des troubles fonctionnels méconnus (Barratt et al., 2017).

Ce contexte souligne l'importance de développer des outils innovants, notamment fondés sur l'intelligence artificielle, pour améliorer la prédiction de formes sévères d'azoospermie à partir de données cliniques et biologiques. Ce projet, réalisé à partir d'un échantillon issu de la population algérienne, vise à renforcer la compréhension des déterminants de l'infertilité masculine dans un contexte socioculturel particulier, tout en proposant un outil d'aide à la décision pour les cliniciens. Notre contribution s'inscrit dans le cadre de l'amélioration du diagnostic de l'infertilité masculine grâce à l'intelligence artificielle, qui constitue le pilier central de notre démarche. Nous proposons une approche innovante visant à exploiter des techniques avancées d'apprentissage automatique afin de prédire la présence ou l'absence de spermatozoïdes dans la biopsie testiculaire à partir de données cliniques, biologiques et hormonales recueillies en Algérie. Cette démarche vise à optimiser la prise en charge des

hommes infertiles en réduisant la nécessité de procédures invasives, tout en tenant compte des particularités socio-culturelles locales.

Le manuscrit est organisé en quatre chapitres. Le premier chapitre traite du rôle des hormones dans la spermatogenèse, base physiologique essentielle pour comprendre les mécanismes de la fertilité masculine. Le deuxième chapitre aborde les différentes causes de l'infertilité masculine, ainsi que les applications actuelles et les limites de l'intelligence artificielle dans ce domaine. Le troisième chapitre décrit le matériel et les méthodes utilisés, notamment la nature des données collectées et le développement du modèle prédictif d'IA. Enfin, le quatrième chapitre présente les résultats obtenus, leur interprétation, ainsi que les perspectives cliniques et sociales de notre travail.

Problématique générale

Comment développer un modèle d'intelligence artificielle fiable et adapté, capable de prédire efficacement la présence ou l'absence de spermatozoïdes dans la biopsie testiculaire chez les hommes infertiles en Algérie, afin d'améliorer le diagnostic et la prise en charge clinique ?

PARTIE 1 : RECHERCHE BIBLIOGRAPHIQUE

CHAPITRE 1 : ROLE DES HORMONES DANS LA SPERMATOGENESE

Introduction

La spermatogenèse est un processus physiologique hautement spécialisé, coordonné par un réseau complexe d'interactions hormonales, cellulaires et moléculaires. Cette séquence dynamique aboutit à la production de spermatozoïdes matures, aptes à la fécondation. Au cœur de ce mécanisme, l'axe hypothalamo-hypophyso-testiculaire joue un rôle régulateur essentiel via la sécrétion des hormones folliculo-stimulante (FSH), lutéinisante (LH) et de la testostérone.

La FSH agit principalement sur les cellules de Sertoli, qui soutiennent la maturation des cellules germinales, tandis que la LH stimule les cellules de Leydig pour la production de testostérone, hormone indispensable à la spermatogenèse et au maintien des caractères sexuels secondaires. Toute perturbation de cet équilibre hormonal, qu'elle soit d'origine centrale ou testiculaire, peut entraîner des anomalies spermatogéniques, conduisant à des troubles de la fertilité masculine tels que l'azoospermie ou l'oligospermie.

L'identification précise des biomarqueurs hormonaux et leur modulation au cours des pathologies testiculaires est donc cruciale pour le diagnostic et le suivi clinique. Ce chapitre s'attache à décrire en détail les rôles respectifs de la FSH, de la LH et de la testostérone dans la spermatogenèse, en soulignant les mécanismes physiopathologiques susceptibles d'être exploités dans le cadre de modèles prédictifs, notamment ceux intégrant des approches d'intelligence artificielle (IA), comme développé dans ce travail.

I.ROLE DES HORMONES DANS LA SPERMATOGENESE

La spermatogenèse débute à la puberté vers l'âge de 12-14 ans sous l'élévation progressive de taux de gonadotrophines (FSH : folliculo-stimulating hormon et LH : luteising hormon) et se poursuit pour toute la vie. C'est un ensemble d'évènements qui se déroulent au sein des tubes séminifères aboutissant aux gamètes mâles haploïdes (les spermatozoïdes). Ce phénomène a une durée de 74 jours et provient d'une manière cyclique et régulière. (Frydman & Poulain, 2023). L'épithélium des tubes séminifères repose sur une lame basale riche en collagène, et comprend deux types cellulaires majeurs : les cellules de Sertoli et les cellules germinales en différenciation (Moore et al., 2016).

Les cellules de Sertoli se développent in utero et dans les premiers mois après la naissance. Elles jouent un rôle essentiel dans la spermatogenèse :

- Mécanique : elles assurent le support structural et forment la barrière hémato-testiculaire, protégeant les cellules germinales du système immunitaire (Nieschlag et al., 2010).
- Biochimique : elles sécrètent protéines (ABP, AMH, Inhibine, Activine), facteurs de croissance, enzymes et nutriments nécessaires à la maturation germinale (Griswold, 2016).
- Phagocytique : elles éliminent les corps résiduels en fin de spermiogénèse (Alberts et al., 2014).
- Contractile : elles favorisent le déplacement des cellules germinales vers la lumière du tube (Smith & Walker, 2014).

La spermatogenèse se divise en trois phases principales :

- Prolifération : les spermatogonies subissent des mitoses successives. À la puberté, sous l'effet de la LH, elles donnent naissance aux spermatocytes I (Moore et al., 2016).
- Méiotique : les spermatocytes I subissent deux divisions successives pour former des spermatides haploïdes (Nieschlag et al., 2010).
- Spermiogenèse : les spermatides se différencient en spermatozoïdes matures par condensation nucléaire, formation de l'acrosome, développement du flagelle et élimination du cytoplasme. La libération finale des spermatozoïdes vers la lumière est appelée spermiation (Tzfira, 2008)

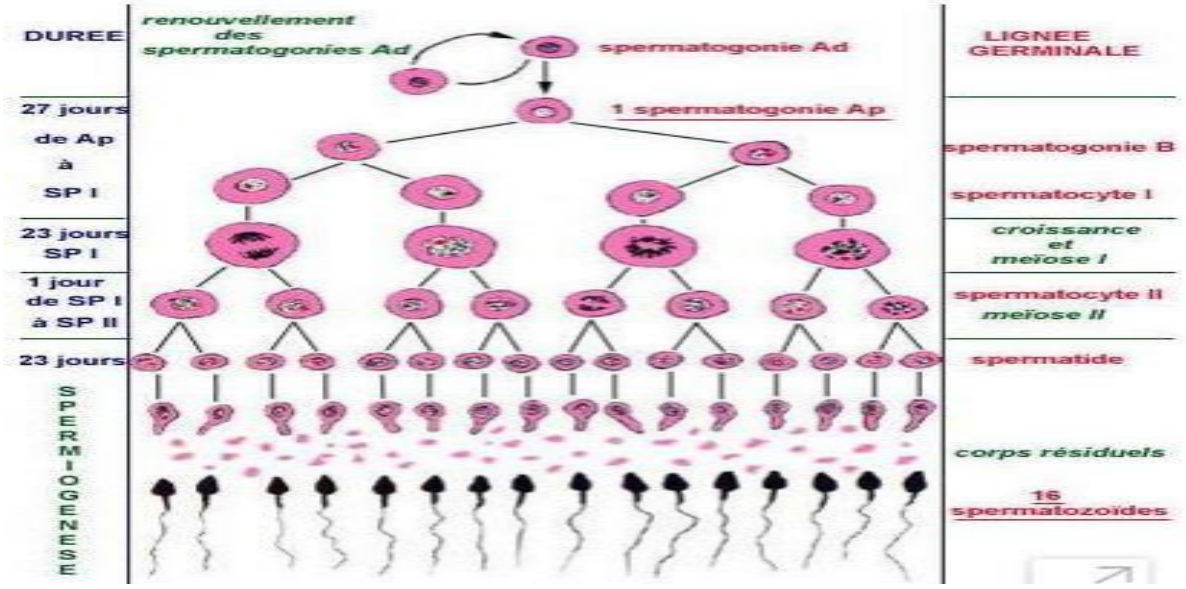

Figure 1 Les étapes de la spermatogenèse.

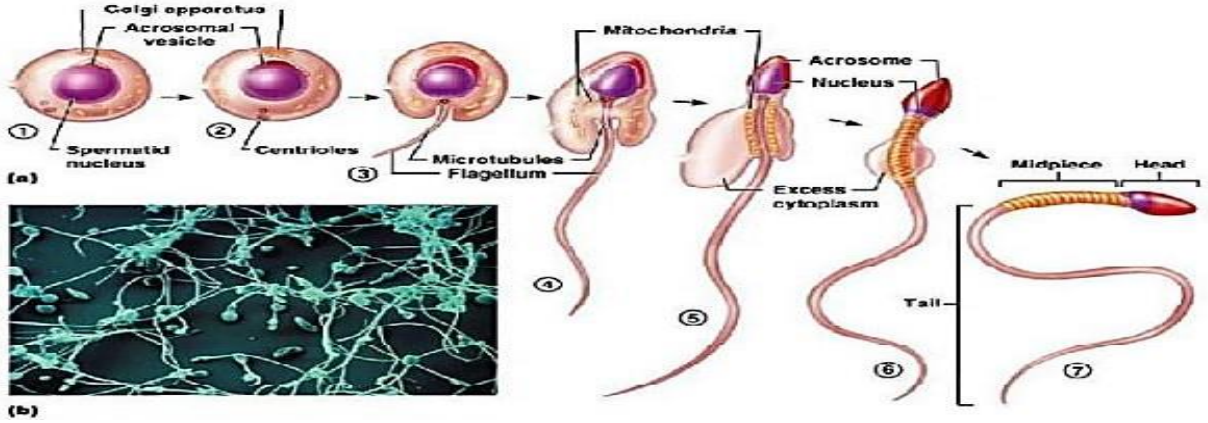

Figure 2 Les étapes de la spermiogénèse.

I.1.Caractéristiques des spermatozoïdes de morphologie normale :

Selon l'OMS un spermatozoïde est dit morphologiquement normal lorsqu'il présente :

- Une tête : qui présente un contour très régulier, ovalaire avec un grand axe mesurant 5µm et un petit axe mesurant 3µm (rapport grand axe/petit axe = 1,66). La longueur et/ou la largeur de la tête peut être légèrement diminuées sans que celle-ci soit pour autant considérée comme anormale. Le rapport possible grand axe/petit axe peut donc fluctuer entre 1,33 et 2cm (s. d.) L'acrosome doit couvrir 40 et 70 % de la surface de la tête, avoir un contour régulier et une texture homogène.

- La pièce intermédiaire : normale peu visible au microscope conventionnel et mesure de 1,5 à 2 fois la longueur de la tête, a un diamètre de 0,6 d 0,8 µm, son grand axe est dans le prolongement du grand axe de la tête, présente un contour régulier, une texture homogène et un reste cytoplasmique de taille minime à son niveau n'est pas considéré comme anormal.

- Enfin, **la pièce principale**, c'est-à-dire le flagelle, mesure environ 45µm (soit environ 10 fois la longueur de la tête), à un diamètre de l'ordre de 0,4 à 0,5 µm, il est développé avec un contour régulier et un aspect homogène

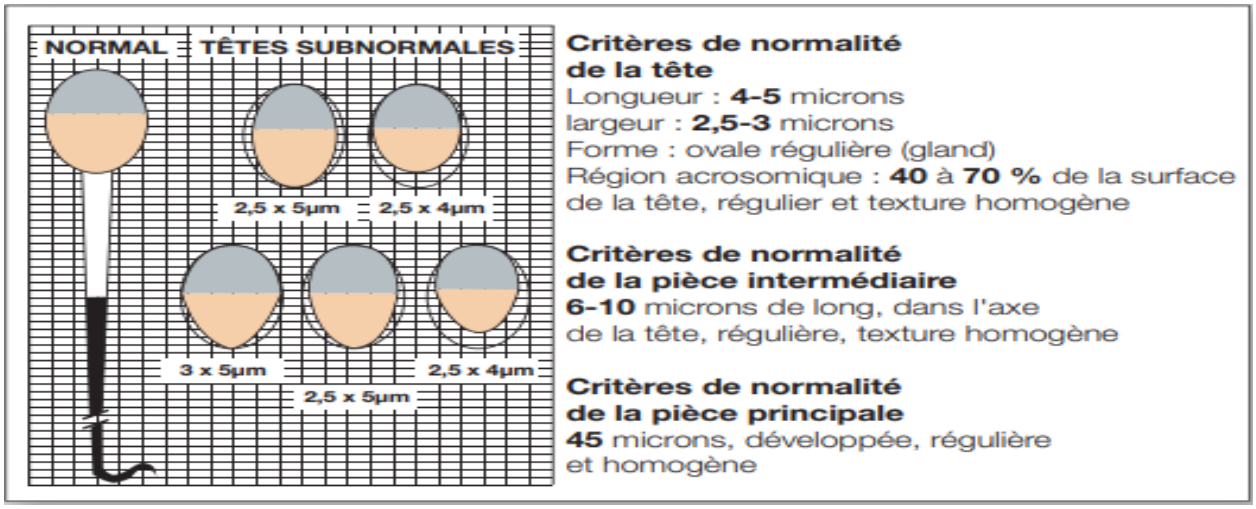

Figure 3 Critères d'un spermatozoïde normal

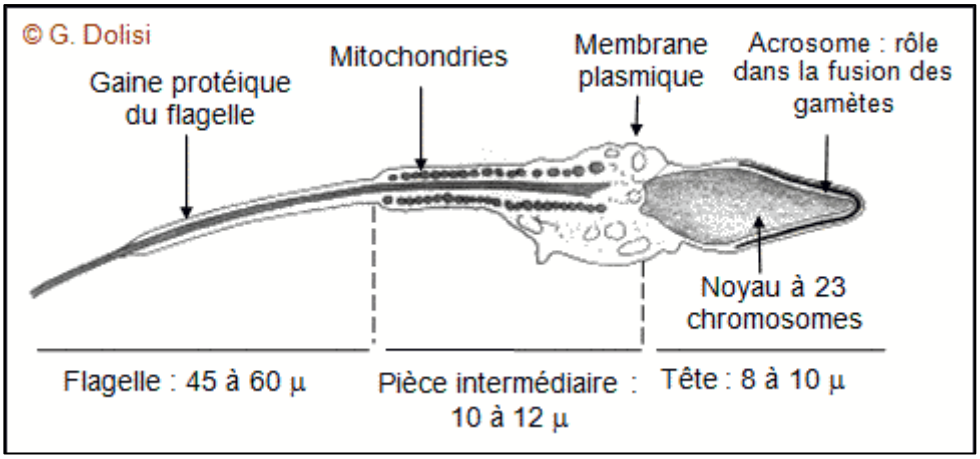

Figure 4 : Spermatozoïde humain au microscope à transmission

I.2.Composition et formation du sperme :

Le sperme est une suspension de spermatozoïdes dans le liquide séminal, issu des sécrétions des glandes génitales mâles (prostate, vésicules séminales). Il contient entre 20 et 100 millions de spermatozoïdes par millilitre, dont au moins 40 % sont mobiles. Le fructose, principal nutriment produit par les vésicules séminales, constitue le carburant énergétique des spermatozoïdes matures. Le liquide séminal renferme également protéines, lipides, prostaglandines (qui facilitent la progression dans les voies génitales féminines en réduisant la viscosité et stimulant un anti-péristaltisme), relaxine et enzymes améliorant la mobilité des spermatozoïdes. Le pH basique (7,2–7,6) neutralise l'acidité de l'urètre et du vagin, optimisant la mobilité. Après l'éjaculation, le sperme coagule puis se liquéfie grâce aux enzymes prostatiques, notamment la (Silber, 2018)

I.3.Régulation de la spermatogenèse

La spermatogenèse résulte d'un équilibre entre mitose, méiose et apoptose, régulé par l'axe hypothalamo-hypophyso-testiculaire. L'hypothalamus sécrète la GnRH, stimulant la libération de FSH

et LH par l'adénohypophyse. La FSH active les cellules de Sertoli, induisant la production de l'ABP qui fixe la testostérone, essentielle au maintien de la spermatogenèse. La LH stimule les cellules de Leydig à produire la testostérone, qui déclenche la spermatogenèse à la puberté et maintient les caractères sexuels secondaires.

Après la naissance, les taux de FSH, LH et testostérone sont élevés, puis diminuent jusqu'à la puberté, où une augmentation de la GnRH relance la production hormonale et la spermatogenèse. La prolactine peut aussi moduler cet axe, bien que son mécanisme soit encore peu clair. Une hyperprolactinémie inhibe l'axe hypothalamo-hypophysaire, réduisant la testostérone (Frydman, 2016)

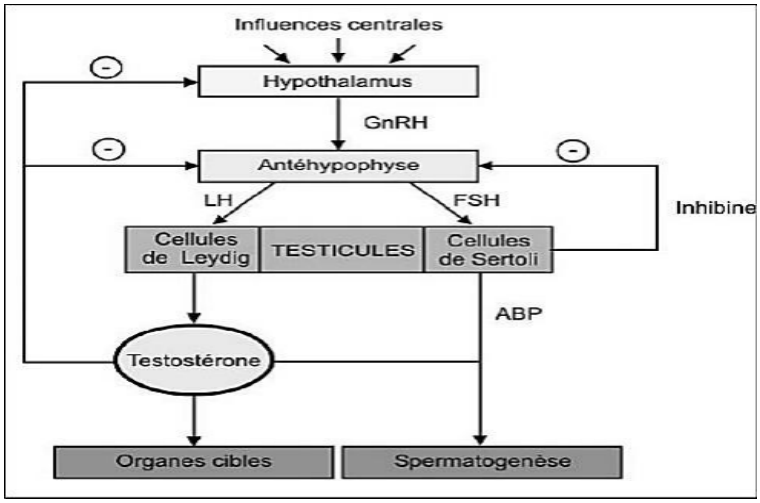

Figure 5: Régulation hormonale de la spermatogenèse.

I.4.Régulation paracrine de la spermatogenèse :

L'activine, sécrétée par les cellules de Sertoli lorsque la production de spermatozoïdes est diminuée, stimule la sécrétion de FSH par l'adénohypophyse, favorisant ainsi la spermatogenèse. Inversement, lorsque la spermatogenèse est active, ces mêmes cellules sécrètent de l'inhibine B, qui exerce un rétrocontrôle négatif sur la FSH. Bien que son rôle chez l'adulte soit encore partiellement élucidé, l'inhibine semble impliquer dans la régulation de la prolifération des cellules de Leydig et de la stéroïdogenèse. Sur le plan clinique, son dosage est pertinent, notamment pour la détection de testicules ectopiques chez les enfants avec gonades non palpables.

CHAPITRE 2 : LES CAUSES DE L'INFERTILITE MASCULINE

Introduction

L'infertilité masculine est une pathologie multifactorielle, résultant d'un ensemble hétérogène de causes qui peuvent être classées selon leur localisation et mécanisme en causes pré-testiculaires, testiculaires, post-testiculaires ou idiopathiques.

Les causes pré-testiculaires regroupent essentiellement les troubles hormonaux d'origine hypothalamo-hypophysaire, entraînant une insuffisance de stimulation testiculaire. Les causes testiculaires incluent des anomalies génétiques (comme les microdélétions du chromosome Y), des pathologies structurales (varicocèle, fibrose testiculaire), ou des dommages acquis (infections, expositions toxiques).

Les causes post-testiculaires, quant à elles, concernent les obstructions des voies spermatiques empêchant l'émission des spermatozoïdes. En dépit des progrès diagnostics, une proportion non négligeable d'infertilité reste classée idiopathique, témoignant de la complexité et de la diversité des facteurs en jeu. Dans ce contexte, l'intégration de données multidimensionnelles (cliniques, biologiques, génétiques) devient indispensable.

L'intelligence artificielle apparaît alors comme un outil innovant capable de traiter ces données hétérogènes et de développer des modèles prédictifs performants. Ce chapitre examine les différentes causes de l'infertilité masculine, en mettant en lumière les limites des méthodes traditionnelles et le potentiel de l'IA pour améliorer le diagnostic et proposer des stratégies personnalisées de prise en charge, conformément à la démarche adoptée dans ce projet de fin d'études.

II.ÉTIOLOGIES DE L'INFERTILITE MASCULINE

La production des spermatozoïdes fonctionnels implique une série de phénomènes physiologiques commençant par la formation des spermatogonies, les cellules de soutien et une différenciation de la gonade primitive en testicule lors de la vie fœtale. Une puberté avec une excrétion androgénique principalement ; la testostérone. Spermatogenèse et transit des spermatozoïdes vers les voies post testiculaires. Formation du sperme et maturation des spermatozoïdes. Toute perturbation pouvant affecter chacune de ces étapes peut être responsable d'infertilité ou de stérilité chez l'homme. L'infertilité masculine peut schématiquement être expliquée par 3 principaux mécanismes, et peut être également inexpliquée, elle est alors dite idiopathique.

- Causes pré-testiculaires : Hypogonadisme hypogonadotrope.
- Causes testiculaires primitives.
- Causes post testiculaires ou obstructives.
- Causes idiopathiques.

II.1Causes pré testiculaires : Hypogonadisme hypogonadotrope

Les causes pré testiculaires sont très rares et touchent l'axe hypothalamo hypophysaire qui est en temps normal responsable de la production des spermatozoïdes et de l'intégrité des fonctions testiculaires endocrines (Guo, 2012). Ceci dit que toute anomalie affectant la première ligne c'est-à-dire les gonadotrophines hypophysaires, peut à la fois être responsable d'un hypogonadisme hypogonadotrope.

L'hypogonadisme hypogonadotrope peut être d'origine :

II.1.1. L'hypogonadisme hypogonadotrope congénital

se définit par un défaut de production de testostérone associé à des taux de gonadotrophines abaissés (FSH et LH). Cliniquement c'est un syndrome qui inclue une absence ou une maturation incomplète sexuelle à l'âge de 18 ans, avec micropénis et/ou cryptorchidie, des taux abaissés de gonadotrophines circulantes et de testostérone. Il peut s'accompagner d'une anosmie ou d'une hyposmie liée à une hypoplasie des bulbes olfactifs caractéristiques du syndrome de Kallmann (Schlosser et al., 2007)

II.1.2. L'hypogonadisme hypogonadotrope acquis

se manifeste par une perte de la libido, une impuissance, une diminution de la pilosité, une oligospermie voire une azoospermie (taux faible ou nul de spermatozoïdes) ou une gynécomastie (Schlosser et al., 2007)

Tableau 1:Causes pré testiculaires de l'infertilité masculine

Type de cause	Particularité	
Cause congénitales	**HHC normosmique (pas de troubles olfactifs) isoles** : déficit restreint a l'axe gonadotrope	Mutations de GNRH1 gène codant pour le GNRH KISS1 gène codant pour la kissepeptine
	Syndrome de KALLMAN : agénésie ou hypoplasie des bulbes olfactifs HHC+anosomie /hyposmie	Mutation de KAL1(ANOS1) liées au chromosome X
	Hypopituitarisme congénitales : diminution des sécrétions hormonales du lobe antérieur de l'hypophyse	Mutation des gènes codant pour les facteurs de transcription impliqués dans le développement de l'hypophyse
Causes acquises	**Tumeurs de la région hypothalamo-hypophysaire :**	Craniopharyngiome. Adenomes hypophysaires dysgerminomes glioma
	Processus infiltratifs hypothalamo-hypophysaires	Hémochromatose juvénile et post transfusionnelle surcharge de fer dans l'organisme . Hypophysite infundibulite

II.2.Causes testiculaires

Les causes testiculaires sont des anomalies qui siègent au niveau des testicules et qui altèrent le déroulement de la spermatogenèse. On peut citer : Les causes congénitales et les causes acquises.

II.3. Causes congénitales

II.3.1. Cryptorchidie

La cryptorchidie est une anomalie congénitale caractérisée par l'absence de descente d'un ou

des deux testicules dans le scrotum. Cette pathologie constitue un facteur majeur d'infertilité masculine, en raison de son impact négatif sur le développement et la maturation des cellules germinales. En effet, la température anormalement élevée à laquelle sont exposés les testicules non descendus perturbe la spermatogenèse. Ils ont montré que l'absence des spermatogonies adultes (Ad) au moment de l'orchidopéxie est fortement corrélée à une infertilité persistante, même après correction chirurgicale. Ainsi, l'intervention précoce est essentielle pour limiter les dommages irréversibles sur la fertilité. Ces résultats soulignent l'importance du dépistage et du traitement rapide de la cryptorchidie dans la prévention de l'infertilité masculine (Hadziselimovic et al., 2007)

II.3.2. Varicocèle

La varicocèle clinique est définie comme des veines anormalement dilatées et tortueuses dans le plexus pampéniforme du cordon spermatique et est classée en fonction des résultats obtenus à la palpation (Hadziselimovic et al., 2007)

La varicocèle se développe habituellement du côté gauche en raison de la prédisposition anatomique de la veine spermatique interne, qui se draine de ce côté à angle droit dans la veine rénale, et est ainsi exposée à une augmentation de la pression veineuse. Elle est à 80% bilatérale. Elle est due à une altération du flux veineux du testicule qui reflue dans la veine spermatique suite à un dysfonctionnement des valves. Le sang contient des substances vasoactives qui induisent une vasoconstriction puis une accumulation des substances toxiques et oxydantes. Ceci réduit la perfusion testiculaire (Singh & Singh, 2017)

Tableau 2:Grades de la varicocèle et leurs caractéristiques cliniques

Grade de varicocèle	Caractéristiques cliniques
Subclinique/ infraclinique	Non palpable Examiné avec l'échographie Doppler et défini par Inversion du flux sanguin veineux lors de la manœuvre de Valsalva où Ectasie de la veine spermatique (>3 mm).
Varicocèle clinique grade 1	Palpable uniquement pendant la manœuvre de Valsalva
Varicocèle clinique de grade 2	Palpable sans la manœuvre de Valsalva
Varicocèle clinique de grade 3	Visible et palpable dans la manœuvre de Valsalva

La varicocèle altère l'histologie des cellules de Leydig, diminuant la sécrétion de la testostérone, la fonction et l'histologie des cellules de Sertoli, diminuant ainsi le support de la spermatogenèse et donc la concentration en spermatozoïdes induisant à une oligozoospermie et augmentant la desquamation des cellules des spermatozoïdes qui se retrouvent sous forme de cellules rondes dans l'éjaculat (Singh & Singh, 2017)

II.3.3. Torsion testiculaire

La torsion testiculaire correspond à la rotation du cordon spermatique, entraînant une interruption brutale de l'apport sanguin au testicule. Cette urgence urologique survient principalement chez l'enfant et l'adolescent et peut rapidement conduire à une ischémie testiculaire irréversible si la prise en charge chirurgicale n'est pas effectuée dans les premières heures. La rapidité du diagnostic et du traitement est essentielle pour préserver la fonction testiculaire et la fertilité future (Ciftci et al., 2008).

II.3.4. Traumatisme testiculaire

Le traumatisme testiculaire résulte d'un choc direct ou d'une compression des testicules, pouvant provoquer des lésions allant du simple hématome à la rupture testiculaire. Bien que rare, ce type de traumatisme peut entraîner des complications sévères, dont une altération de la fonction reproductive. Le diagnostic repose sur l'examen clinique et l'échographie doppler, qui guide la décision thérapeutique entre traitement conservateur et intervention chirurgicale (Van Den Bergh, 2011)

II.4. Causes iatrogènes

Les infertilités masculines acquises peuvent être d'origine médicamenteuse. Néanmoins, le mécanisme exact reste controversé. Plusieurs médicaments, tels que certains agents chimiothérapeutiques , les glucocorticoïdes, et certains psychotropes, peuvent altérer la spermatogenèse, la production hormonale ou la qualité du sperme. Ces effets peuvent être temporaires ou permanents selon la nature et la durée du traitement. Cependant, la complexité des mécanismes sous-jacents et la variabilité interindividuelle compliquent la compréhension complète de ces effets ((Razzak, 2012)

Tableau 3:Causes iatrogènes de l'infertilité masculine et leurs mécanismes

MECANISME	MEDICAMENTS
Inhibition du transport Des Spermatozoïdes	Antihypertenseurs et psychotropes.
Suppression de la spermatogenèse	Agents cytostatiques
	Hormones et stéroïdes : androgènes, anti- androgènes, œstrogènes, progestatifs, glucocorticoïdes, anabolisants, cimétidine, spironolactine, digoxine, kétoconazole.
	Psychotropes, antiépileptiques, antiémétiques Analgésiques, certains antibiotiques, chimiothérapies, antihélmintiques
Altération de la fonction des spermatozoïdes	Anticalciques (mobilité des spermatozoïdes et liaison des spermatozoïdes à l'ovocyte)
	Antiépileptiques (mobilité des spermatozoïdes)
	Sulphazalasine (mobilité et numération des spermatozoides)
	Antibiotiques (mobilité des spermatozoïdes)
	Amantadine et colchicine (interaction spermatozoïde-ovocyte).
	Psychotropes et bétabloquants (mobilité)

II.5. Causes infectieuses

II.5.1. Oreillons

Les oreillons ou parotidite ourlienne sont une maladie virale due au genre rubulavirus atteignant principalement les glandes salivaires et le tissu nerveux. Elle survient habituellement chez l'enfant de plus de deux ans mais peut atteindre l'adulte et est contagieuse par des suspensions de gouttelettes de salive. Parmi ses complications, on trouve l'orchite qui est une inflammation des testicules qui survient chez 50% des malades et qui provoque une atrophie des testicules pouvant aller jusqu'à une diminution

de la fertilité et rarement la stérilité. A noter que l'orchite est observée dans d'autres cas d'infections virales telles que celles à l'herpès virus. (Singh & Singh, 2017)

II.5.2.MST et infertilité masculine

Les maladies sexuellement transmissibles sont des infections dues à des bactéries, virus ou parasites transmis par contact vénérien. Parmi les principales bactéries incriminées, on trouve Chlamydia trachomatis (chlamydiose), Neisseria gonorrhoeae (gonorrhée), Treponema pallidum (syphilis), Mycoplasma et Ureaplasma espèces. Les infections virales MST sont dues aux virus HSV, HIV, VPH, et les virus de l'hépatite B et C. Le parasite principal incriminé dans les MST est Trichomonas vaginalis (« Report on Optimal Evaluation of the Infertile Male », 2006)

II.5.3.Infections bactériennes

- ***Chlamydia trachomatis*** est une IST fréquente, responsable d'urétrites, épididymites et prostato-vésiculites chez l'homme. Son effet sur le sperme est controversé, mais elle altère le tractus génital, réduisant la spermatogenèse et la perméabilité des voies spermatiques (Pacey & Eley, 2004)
- **Infection à *Neisseria gonorrhoeae*** (gonocoque) provoque la blennorragie, une urétrite courante. L'inflammation qu'elle induit réduit la fertilité en affectant la motilité et la morphologie des spermatozoïdes via une élévation des leucocytes dans le sperme (« Report on Optimal Evaluation of the Infertile Male », 2006)

II.5.4.Infections parasitaires

- ***Trichomonas vaginalis :*** C'est un parasite flagellé qui touche tant l'homme que la femme. Les hommes infectés montrent une inflammation, une irritation et une urétrite. Des études récentes ont montré que le *T.vaginalis* produit des substances toxiques qui peuvent réduire la motilité, la vitalité ainsi que la fonction des spermatozoïdes et peuvent même les détruire directement. (Schill et al., 2008)

II.6.Causes post testiculaires

Tout obstacle atteignant ces canaux est responsable d'une anomalie de l'éjaculation

II.6.1.Obstruction post-spermatogenèses

L'obstruction post-spermatogénétique correspond à un blocage des voies d'évacuation des spermatozoïdes, notamment au niveau de l'épididyme, du canal déférent ou des conduits éjaculateurs, alors que la spermatogenèse testiculaire reste normale. Cette obstruction est une cause majeure d'azoospermie obstructive, entraînant une infertilité masculine. Les causes peuvent être congénitales (comme l'absence bilatérale des canaux déférents) ou acquises (infections, traumatismes, interventions chirurgicales). Le diagnostic repose sur l'examen clinique, le spermogramme et l'imagerie. Le traitement inclut la chirurgie réparatrice ou la récupération des spermatozoïdes pour la procréation assistée (Kahn et al., 1975).

II.6.2. Anomalie des canaux déférents

La cause majeure d'obstacle post testiculaire est l'agénésie bilatérale des canaux déférents (ABCD) qui est une maladie génétique autosomique récessive fréquente, liée à des mutations bi-alléliques du gène CFTR qui est responsable de la mucoviscidose. La différence entre les deux maladies c'est que la mucoviscidose est due à deux mutations sévères au niveau du gène CFTR tandis que l'ABCD est due à une mutation sévère associée à une mutation mineure de CFTR. Etant donné la très grande fréquence des porteurs asymptomatiques (près de 1/30) dans la population générale française, une anomalie de CFTR chez la partenaire d'un patient avec ABCD sera systématiquement recherchée dans le cadre du conseil génétique (De Braekeleer & Férec, 1996)

II.7. Causes immunologiques

La constitution chromosomique des spermatozoïdes est différente de celle des cellules somatiques, ce qui fait que les antigènes spermatiques sont considérés comme étrangers par les cellules du système immunitaire. Cependant, lors de l'apparition des spermatozoïdes à la puberté, un isolement de ces cellules est obtenu par la barrière hémato-testiculaire formée par Les jonctions serrées entre les cellules de Sertoli assurent une barrière immunitaire au niveau testiculaire. La rupture de cette barrière, un défaut d'immunosuppression ou un traumatisme peuvent entraîner une réponse immunitaire contre les spermatozoïdes. Parmi les causes impliquées : vasectomie, obstruction, infections, traumatismes, varicocèle, ou encore intolérance aux métaux lourds. Les anticorps anti-spermatozoïdes (AAS) altèrent la fertilité en réduisant la capacité fécondante et en induisant un stress oxydatif destructeur. En l'absence de cause infectieuse ou obstructive, un traitement immunosuppresseur ou la PMA est envisagé. (Singh & Singh, 2017)

II.8. Causes génétiques

II.8.1. Syndrome de Klinfelter

C'est une anomalie des chromosomes touchant particulièrement les gonosomes, dans ce cas le nombre de chromosomes X est augmenté d'au moins un. Ce chromosome supplémentaire affecte directement les cellules souches spermatiques altérant leur renouvellement et induisant une apoptose des spermatogonies, ce qui provoque une interruption précoce de la spermatogenèse au stade pré-méiotique. Ce syndrome est une cause majeure de l'infertilité masculine puisqu'il est retrouvé chez environ 15% des hommes azoospermes.

Sur le plan phénotypique, le SK est associé à une hypotrophie majeure des testicules et parfois à un hypogonadisme avec gynécomastie(Gueye et al., 1999).

II.8.2. Microdélétions du bras long du chromosome Y

C'est l'étiologie génétique la plus fréquente puisqu'elle est observée chez 10% des hommes infertiles ayant une azoospermie obstructive, et chez 5% des hommes infertiles ayant une oligospermie sévère. Ces microdélétions sont des pertes plus ou moins importante dans la région AZF du chromosome Y. Cette région est composée de 3 locus : AZFa, AZFb, AZFc où se trouvent de nombreuses séquences d'ADN et où surviennent ces délétions. Ceci entraine une grosse perte de gènes qui interviennent dans la production des spermatozoïdes. La région la plus touchée est celle du locus AZFc qui est observée chez 1 homme/2300 et est responsable d'une perte importante de 4 copies du gène DAZ (Delete in azoospermia) qui sont nécessaires, dans leur ensemble à la spermatogenèse (Schlosser et al., 2007)

II.8.3. Mutation de CFTR

La mutation du gène CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) est une cause génétique majeure d'infertilité masculine, notamment en raison de son association avec l'absence congénitale bilatérale des canaux déférents (CBAVD). Cette condition entraîne une obstruction post-spermatogénétique, empêchant le passage des spermatozoïdes malgré une spermatogenèse normale. Les mutations les plus courantes associées à la CBAVD sont ΔF508, R117H et IVS8-5T. Il est essentiel de réaliser un dépistage génétique chez les hommes présentant une azoospermie obstructive pour identifier ces mutations et évaluer le risque de transmission à la descendance (de Souza et al., 2018).

II.9. Causes Environnementales et Toxiques

II.9.1. Le Tabac

De nombreuses études montrent que le tabagisme nuit à la quantité et à la qualité des spermatozoïdes, tant chez les hommes fertiles qu'infertiles. Par exemple, Kunzle et al. (2003) ont observé chez des hommes infertiles une baisse de 17,5 % du nombre de spermatozoïdes et de 16,6 % de leur motilité chez les fumeurs par rapport aux non-fumeurs. Bien qu'une autre étude n'ait pas trouvé de différence dans les paramètres classiques, elle a révélé une augmentation significative de la fragmentation de l'ADN spermatique chez les fumeurs ((Frankle, 1976)). Cette fragmentation est un indicateur crucial, car elle est liée à un risque accru d'avortements spontanés et à un moindre succès des traitements de fertilité (Bland et al., 1976).

II.9.2. L'Alcool

Une étude indique que 17 % des hommes en clinique de fertilité consomment de l'alcool (Homan et al., 2012). Chez les couples infertiles, l'alcool réduit aussi les chances de réussite des traitements de

procréation, avec un risque d'avortement spontané doublé chez les femmes dont le partenaire ((« V.I. Gavrilov », 1975)).

II.9.3. Exposition aux produits chimiques ou métaux lourds

L'exposition aux produits chimiques toxiques et aux métaux lourds constitue un facteur important d'infertilité masculine acquise. Ces substances, souvent présentes dans l'environnement industriel, agricole ou domestique, peuvent perturber la spermatogenèse et altérer la qualité du sperme. Parmi les métaux lourds, le plomb, le cadmium, le mercure et l'arsenic sont particulièrement reconnus pour leurs effets toxiques sur le système reproducteur masculin, notamment par induction de stress oxydatif, inflammation testiculaire et dysfonctionnement hormonal. De même, certains solvants organiques, pesticides, et hydrocarbures aromatiques polycycliques (HAP) sont impliqués dans la réduction du nombre et de la mobilité des spermatozoïdes, ainsi que dans des anomalies morphologiques. Ces effets délétères peuvent entraîner une infertilité temporaire ou permanente selon la durée et le niveau d'exposition, rendant indispensable la prévention et le suivi médical des populations à risque (Durbin, 1975).

II.10. Causes idiopathiques

Les causes idiopathiques de l'infertilité masculine désignent les cas où, malgré un bilan médical complet, aucune étiologie identifiable n'est détectée. Cette situation concerne environ 30 à 40 % des hommes infertiles. Les mécanismes sous-jacents restent mal compris, mais plusieurs pistes sont explorées. Des anomalies subcliniques, telles que des dysfonctionnements des cellules de Sertoli ou des altérations épigénétiques, pourraient perturber la spermatogenèse. De plus, des facteurs environnementaux, comme l'exposition à des polluants ou à des espèces réactives de l'oxygène, peuvent induire des dommages à l'ADN spermatique, affectant ainsi la qualité du sperme et la fertilité. Le diagnostic repose sur l'élimination des causes connues, et la prise en charge inclut des traitements antioxydants, hormonaux ou des techniques de procréation assistée. Cependant, la compréhension de ces mécanismes nécessite encore des recherches approfondies (DeMeester & Johnson, 1975).

III. L'IA dans les traitements des données de santé

L'intelligence artificielle (IA) s'impose aujourd'hui comme un levier majeur d'innovation dans le domaine médical. Grâce à sa capacité à traiter d'immenses volumes de données hétérogènes (imagerie, analyses biologiques, génomique, dossiers médicaux électroniques), l'IA offre des perspectives prometteuses pour améliorer le diagnostic, prédire l'évolution des maladies, personnaliser les traitements, et optimiser la gestion des soins.

Les algorithmes d'apprentissage automatique (machine learning) et d'apprentissage profond (deep learning) permettent d'extraire des schémas invisibles à l'œil humain. Ces outils sont déjà utilisés dans plusieurs spécialités :

- **En radiologie**, les systèmes IA surpassent parfois les radiologues dans le dépistage du cancer du sein via mammographie (McKinney et al., 2020), mais aussi dans l'analyse automatisée des scanners pulmonaires ou cérébraux (nodule pulmonaire, AVC).
- **En cardiologie**, Attia et al. (2019) ont développé un modèle capable de prédire la fibrillation auriculaire à partir d'un ECG apparemment normal.
- **En dermatologie**, les réseaux de neurones convolutifs atteignent une précision comparable à celle des experts pour détecter des mélanomes ou carcinomes cutanés (Esteva et al., 2017).
- **En ophtalmologie**, l'IA est utilisée pour le dépistage automatisé de la rétinopathie diabétique à partir de simples photographies du fond d'œil, notamment dans les zones à faible accès médical.

Par ailleurs, l'IA joue un rôle croissant dans la **médecine prédictive** (prédiction du diabète, rechute de cancer, réponse aux traitements), ainsi que dans la **médecine personnalisée**, en suggérant des thérapies ciblées adaptées au profil génétique du patient.

Ces avancées sont rendues possibles par la disponibilité croissante de bases de données médicales massives (big data) et par les progrès des capacités de calcul. Toutefois, ces technologies soulèvent aussi des défis éthiques, réglementaires et humains, notamment en matière de confidentialité, de transparence des algorithmes, et de responsabilité médicale.

III.1. Les modèles d'IA en santé

L'intelligence artificielle (IA) joue un rôle de plus en plus central dans la transformation des pratiques médicales modernes. Grâce à sa capacité à analyser de grandes quantités de données médicales complexes (dossiers électroniques, imagerie, analyses biologiques, génomique), l'IA permet d'améliorer le dépistage, le diagnostic, la prédiction de maladies, ainsi que la personnalisation des traitements.

Les approches d'IA les plus couramment utilisées en médecine reposent sur deux grandes catégories :

- **L'apprentissage automatique (machine learning)** : l'algorithme apprend à partir de données étiquetées (apprentissage supervisé) ou non étiquetées (apprentissage non supervisé), afin d'identifier des modèles, prédire des résultats cliniques ou segmenter des groupes de patients.

- **L’apprentissage profond (deep learning)** : forme avancée de machine Learning utilisant des réseaux de neurones artificiels multicouches, particulièrement efficace dans l’analyse d’images, de signaux ou de texte non structuré.

Par exemple, en radiologie, des réseaux de neurones convolutifs (CNN) ont montré des performances équivalentes, voire supérieures, à celles des radiologues dans le dépistage du cancer du sein à partir de mammographies (McKinney et al., 2020). En cardiologie, un modèle d’IA a permis de prédire la fibrillation auriculaire à partir d’ECG normaux, démontrant une capacité de détection précoce inédite (Attia et al., 2019).

En dermatologie, l’IA peut classer des lésions cutanées avec un niveau de précision équivalent à celui de dermatologues expérimentés (Esteva et al., 2017).

Ces technologies s’appuient sur des architectures mathématiques complexes, mais leur objectif reste simple : extraire une information exploitable pour aider à la décision médicale, réduire les erreurs, et accélérer les processus cliniques.

III.2. Principe de ces modèles

Le principe repose sur l’entraînement d’un algorithme à partir d’exemples annotés. Par exemple, dans un modèle de classification, on fournit des données d’entrée (comme des examens biologiques) associées à un résultat (ex. : présence ou non d’une maladie). L’algorithme apprend à établir une fonction de prédiction applicable à de nouveaux cas. En santé, ces modèles sont utiles pour détecter des pathologies, prédire des réponses à un traitement ou évaluer des risques (Shickel et al., 2018)).

III.3. Utilisation de quelques exemples

Un exemple concret est l’usage de l’IA dans le dépistage du cancer du sein à partir de mammographies. Des modèles développés par Google Health ont montré des performances comparables, voire supérieures, à celles de radiologues humains (McKinney et al., 2020). En urologie, certains algorithmes permettent de prédire la fertilité masculine à partir de paramètres hormonaux, comme cela est exploré dans des travaux récents sur l’apprentissage automatique appliqué à l’andrologie (Chen et al., 2022).

De plus, dans la prédiction du diabète, des modèles comme les forêts aléatoires ou les réseaux de neurones permettent d’identifier précocement les personnes à risque à partir de données simples comme la glycémie, l’âge ou l’IMC ((Ehrhart et al., 1975)). Ces outils constituent de véritables aides à la décision médicale

PARTIE 2 : MATÉRIEL ET MÉTHODES

I. MATÉRIEL

Dans le cadre de cette étude, nous avons adopté une démarche basée sur l'apprentissage automatique (machine learning) supervisé afin de modéliser les facteurs cliniques et biologiques associés à l'infertilité masculine. L'objectif était de construire un modèle prédictif capable d'identifier, à partir de données patient, les cas probables d'azoospermie sévère.

La méthode générale comprend plusieurs étapes clés : le nettoyage des données, leur transformation, l'analyse exploratoire, la modélisation via des algorithmes d'apprentissage, puis l'évaluation des performances. Pour cela, nous avons utilisé un jeu de données cliniques, des outils open-source (Python, Scikit-learn, etc.), et une infrastructure informatique de type personnel.

I.1. Données cliniques

Le fichier principal utilisé dans ce travail est nommé azoospermia_bensaada.csv. Il a été obtenu à partir des dossiers médicaux anonymisés de patients suivis au centre de procréation médicalement assistée de la clinique AVICENNE (Constantine, Algérie), dans le respect des règles éthiques de confidentialité.

Ce fichier est au format CSV (Comma-Separated Values) , largement utilisé en analyse de données biomédicales (Voir le Tableau 1). Il contient les informations cliniques, biologiques et anatomopathologiques de patients suspectés ou diagnostiqués avec une infertilité masculine. La variable cible, intitulée RESULTAT_BT, représente le résultat de la biopsie testiculaire (0 = azoospermie sévère, 1 = présence de spermatozoïdes).

L'objectif du modèle est de prédire cette variable binaire (RESULTAT_BT) à partir des autres données disponibles, permettant ainsi de classer chaque patient dans l'une des deux catégories diagnostiques.

Le fichier comprend les variables suivantes :

- Variables biologiques : FSH, LH, TESTO (taux hormonaux),
- Facteurs cliniques antécédents médicaux : AUTRE_MALADIE, VARICOCELE,
- Résultats d'imagerie et histologie : ECHO_Testuculaire (TG, TD)
- FIBROSE, HYPOSITIFPERM, Sertoli Cell Only,
- Antécédents de maturation : ARRET_DE_M,
- Variable cible : RESULTAT_BT (diagnostic final).

Les données ont fait l'objet d'un prétraitement rigoureux : suppression des doublons, gestion des valeurs manquantes, codage des variables catégorielles, et standardisation des variables numériques

pour une meilleure performance des modèles.

Tableau 4: Données utilisées pour l'apprentissage

	AGE	AUTRE_MALADIE	VARICOCELE	FSH	LH	TESTO	ECHO_TD	TG	HYPOSITIFPERM	ARRET_DE_M	SCO	FIBROSE	RESULTAT_BT
0	29	0	1	15.00	7.50	8.90	1	1	0	1	1	0	0
1	41	1	1	33.12	15.20	2.68	1	2	0	1	1	0	1
2	38	0	1	1.73	1.32	9.13	1	1	0	1	1	1	0
3	31	0	0	11.88	5.51	9.58	0	0	0	1	1	0	1
4	34	0	0	1.90	0.55	1.32	2	2	0	1	1	1	1

I.2. Configuration de la machine

La machine utilisée est un simple ordinateur, voir caractéristiques détaillées dans le tableau 05 :

Tableau 5 Tableau représentant la configuration du matériel informatique utilisé lors de l'apprentissage

Ordinateur	Caractéristiques
Processeur	Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz 2.11 GHz
Mémoire installée RAM	8,00 Go
Stockage	SSD 512,00 Go
Système d'exploitation	Windows 11 professionnel
Type de système	Système d'exploitation 64 bits
Version du système	24H2 | 26100.3775

I.4. Outils et bibliothèques

I.4.1. Environnement de travail

- **Python**

  Python est un langage de programmation open source, interprété et orienté objet, largement utilisé en data science pour sa simplicité et sa lisibilité. Il permet de développer rapidement des applications complexes, tout en facilitant le traitement, l'analyse et la visualisation de données.

- **Anaconda**

  Anaconda est une distribution libre de Python et R, conçue pour simplifier l'installation des bibliothèques et la gestion d'environnements de développement. Elle est particulièrement adaptée aux projets de science des données et de machine learning.

- **Jupyter Notebook**

  Jupyter Notebook est une application web interactive permettant de combiner du code, des visualisations et du texte explicatif dans un même document. Elle est idéale pour l'analyse exploratoire, la modélisation, la visualisation et la documentation simultanée du travail.

I.4.2. Bibliothèques Python utilisées

- **NumPy**

NumPy est une bibliothèque open source fondée en 2005 pour effectuer des calculs numériques efficaces en Python. Elle est essentielle pour la manipulation de tableaux multidimensionnels et fournit un ensemble riche de fonctions mathématiques (“NumPy,” 2020).

- **Pandas**

Pandas est une bibliothèque conçue pour la manipulation et l’analyse des données. Elle propose des structures de données puissantes, telles que les DataFrames, facilitant la gestion et l’analyse de grands ensembles de données (“pandas - Python Data Analysis Library,” 2020).

- **Matplotlib**

Matplotlib est une bibliothèque Python dédiée à la création de graphiques statiques, animés et interactifs. Elle est souvent utilisée en complément de NumPy et SciPy pour la visualisation scientifique des données (“Matplotlib,” 2020).

- **Seaborn**

Seaborn est une bibliothèque de visualisation statistique basée sur Matplotlib. Elle offre une interface plus simple et des graphiques plus esthétiques, facilitant l’analyse exploratoire des données.

- **Scikit-learn (sklearn)**

Scikit-learn est un module Python qui intègre des algorithmes classiques de machine learning dans l’écosystème scientifique Python (numpy, scipy, matplotlib). Il fournit des solutions simples et efficaces pour la régression (linéaire, logistique), la classification (K-Nearest Neighbours), le clustering (K-Means), ainsi que pour le prétraitement des données (normalisation Min-Max, etc.). Il est conçu pour être accessible, réutilisable et polyvalent (“scikit-learn: machine learning in Python,” 2020).

- **TensorFlow**

TensorFlow est une plateforme open source dédiée au machine learning et au deep learning. Elle propose un ensemble complet d’outils, de bibliothèques et de ressources permettant aux chercheurs et développeurs de créer, entraîner et déployer des modèles d’intelligence artificielle de manière efficace (“TensorFlow,” 2020).

Le tableau suivant (Tableau 3) représente les différents outils et bibliothèques avec les versions utilisées pour notre travail.

Tableau 6 *Caractéristiques des différents outils informatiques utilisés*

Outils / bibliothèques	Versions

Anaconda	**2023.11**
Jupyter Notebook	**6.5**
Matplotlib	**3.7**
Numpy	**1.25**
Pandas	**2.0**
scikit-learn	**1.3**
Seaborn	**0.13**
TensorFlow	**2.13**
Python	**3.11**

II.MÉTHODES

Cette partie décrit les méthodes utilisées afin d'aboutir à l'objectif de l'approche.

II.1. Description globale

L'objectif du modèle est de prédire cette variable binaire (RESULTAT_BT) à partir des autres données disponibles, permettant ainsi de classer chaque patient dans l'une des deux catégories diagnostiques.

Dans cette section, nous présentons l'approche globale, qui est présentée également dans la (Figure 3) :

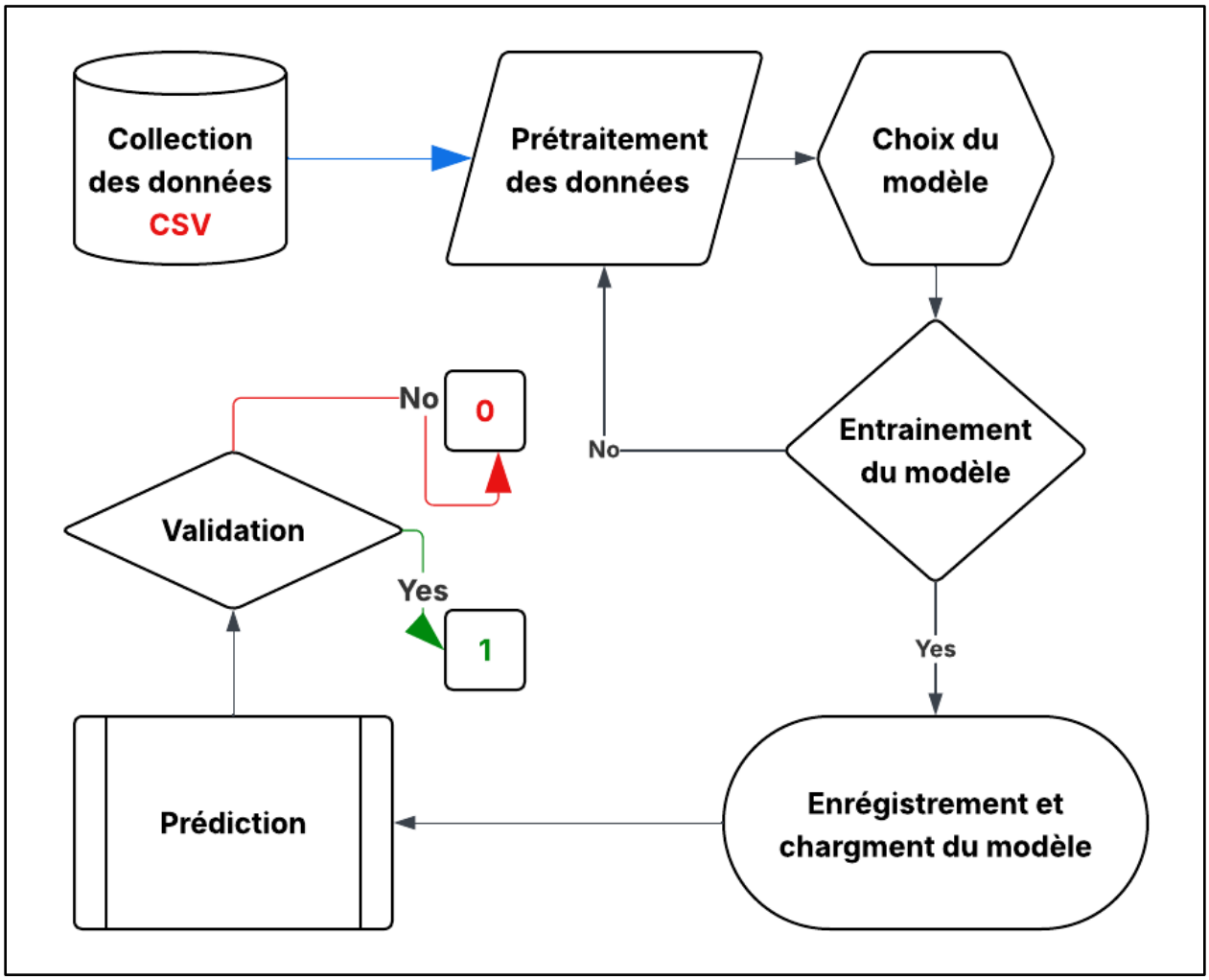

Figure 6 Processus de l'approche globale

II.1.1. Prétraitement des données

Les données utilisées proviennent d'un fichier CSV nommé azoospermia_bensaada.csv, contenant différentes variables biologiques et cliniques en rapport avec l'azoospermie. Afin d'équilibrer les classes cibles (0 pour azoospermie sévère, 1 pour non azoospermie), un suréchantillonnage a été appliqué sur la classe minoritaire pour éviter un biais lors de l'apprentissage. Plusieurs ensembles de variables explicatives ont été définis, correspondant à différentes combinaisons de caractéristiques cliniques. Les données ont ensuite été normalisées par standardisation (moyenne nulle, variance unitaire) grâce à un StandardScaler, ce qui facilite la convergence et la robustesse des modèles d'apprentissage automatique.

II.1.2. Apprentissage

Pour modéliser la prédiction du résultat biologique (RESULTAT_BT), un classifieur Random Forest a été utilisé. Ce modèle d'ensemble construit plusieurs arbres de décision indépendants puis agrège leurs prédictions pour améliorer la robustesse et la performance. L'entraînement a été réalisé via une validation croisée à 5 plis stratifiés, afin d'évaluer la capacité de généralisation du modèle tout en conservant la proportion des classes. Les métriques calculées comprennent l'accuracy, la précision, le rappel et le F1-score, fournissant une évaluation complète de la qualité des prédictions. Chaque sous-ensemble de variables a donné lieu à un modèle spécifique, entraîné sur l'ensemble des données après normalisation. Le modèle intègre également une pondération des classes pour compenser le déséquilibre initial.

II.1.3. Prédiction

Une fois les modèles entraînés, ils ont été utilisés pour prédire la classe sur un jeu de nouvelles données brutes, préalablement normalisées avec les mêmes paramètres que ceux utilisés en apprentissage. Les prédictions fournissent non seulement la classe assignée (0 ou 1) mais aussi la confiance associée à cette prédiction, exprimée en pourcentage. Une comparaison a été réalisée entre les prédictions et les résultats réels connus, mettant en évidence la performance du modèle sur des cas concrets. Enfin, une matrice de confusion visuelle a été générée pour représenter les erreurs et succès des prédictions, facilitant l'interprétation des performances du modèle. Cette démarche complète permet d'utiliser efficacement les modèles pour des diagnostics automatisés sur des patients nouveaux.

II.2. Description détaillée des méthodes

II.2.1. Appel des bibliothèques

Ce script constitue une étape clé dans la mise en œuvre d'un modèle de classification supervisée dans le cadre d'un projet de mémoire de fin d'étude. L'approche repose sur l'utilisation de l'algorithme **Random Forest (forêt aléatoire)**, reconnu pour sa robustesse face aux données bruitées et multivariées, notamment dans le domaine biomédical (Voir figure 4).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import StratifiedKFold, cross_validate
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (
    confusion_matrix, classification_report, accuracy_score,
    roc_curve, auc
)
```

Figure 7 Appel des bibliothèques et préparation du modèle

Le code met en place les éléments suivants :

- **Préparation des données** : importation et manipulation via les bibliothèques `pandas` et `NumPy`.
- **Normalisation** des variables explicatives à l'aide de `StandardScaler` pour optimiser l'apprentissage.
- **Stratified K-Fold Cross-Validation** : une technique rigoureuse de validation croisée qui garantit une répartition équilibrée des classes à chaque pli, améliorant ainsi la robustesse de l'évaluation.
- **Entraînement d'un modèle Random Forest** via `scikit-learn`, avec évaluation sur différents plis de validation.
- **Mesures de performance** : rapport de classification, précision globale (`accuracy`), matrice de confusion, courbe ROC et **aire sous la courbe (AUC)**.

- **Visualisation des résultats** : utilisation de `matplotlib` et `seaborn` pour illustrer les performances du modèle de manière claire et interprétable.

Ce script s'intègre dans une démarche de modélisation prédictive appliquée aux données médicales, avec pour objectif l'aide à la décision clinique, par exemple dans le diagnostic de l'infertilité masculine.

II.2.3. Acquisition des données et exploration préliminaire

```
# Chargement des données
df = pd.read_csv("azoospermia_bensaada.csv")

features = ['AUTRE_MALADIE', 'VARICOCELE', 'FSH', 'LH', 'TESTO',
            'ECHO_TD', 'TG', 'HYPOSITIFPERM', 'FIBROSE']
target = 'RESULTAT_BT'

# Analyse exploratoire
plt.figure(figsize=(8,5))
sns.boxplot(x=target, y='AGE', data=df)
plt.title("Distribution de l'âge selon RESULTAT_BT")
plt.xlabel("RESULTAT_BT (0=azoospermie sévère, 1=non azoospermie)")
plt.ylabel("Âge")
plt.show()
```

Figure 8: Acquisition des données et exploration préliminaire

- **Chargement des données cliniques et analyse descriptive initiale**

Cette section initie l'analyse prédictive par l'importation d'un dataset clinique spécialisé ("azoospermia_bensaada.csv") via pandas, garantissant l'intégrité des données biomédicales. La sélection de neuf variables explicatives suit une approche multidimensionnelle intégrant l'axe clinique (AUTRE_MALADIE, VARICOCELE, HYPOSITIFPERM, FIBROSE), l'axe

endocrinologique (FSH, LH, TESTO) constituant le triptyque hormonal de référence pour l'évaluation de l'axe hypothalamo-hypophyso-gonadique, et l'axe morphologique (ECHO_TD, TG) représentant les données d'imagerie. La variable cible RESULTAT_BT correspond à l'outcome d'une biopsie testiculaire, procédure diagnostique de référence différenciant l'azoospermie obstructive de la forme sécrétoire, cette variable binaire (0 = azoospermie sévère, 1 = non azoospermie) déterminant directement les possibilités thérapeutiques en assistance médicale à la procréation. L'exploration descriptive utilise un boxplot pour analyser la distribution de l'âge selon le résultat diagnostique, permettant d'identifier d'éventuelles associations entre l'âge et l'outcome clinique, l'âge constituant un facteur pronostique reconnu en andrologie pour le counseling thérapeutique des hommes infertiles.

II.2.4. Analyse des distributions hormonales et matrice de corrélation

```
vars_num = ['FSH', 'LH', 'TESTO']
for var in vars_num:
    plt.figure(figsize=(8,4))
    sns.histplot(data=df, x=var, hue=target, kde=True, stat="density", common_norm=False)
    plt.title(f"Distribution de {var} selon RESULTAT_BT")
    plt.xlabel(var)
    plt.ylabel("Densité")
    plt.show()

plt.figure(figsize=(10,8))
corr = df[features + [target]].corr()
sns.heatmap(corr, annot=True, fmt=".2f", cmap='coolwarm', center=0)
plt.title("Matrice de corrélation des variables")
plt.show()
```

Figure 9:Analyse des distributions hormonales et matrice de corrélation

– **Exploration multivariée des profils endocriniens et analyse corrélationnelle**

Cette section approfondit l'analyse exploratoire par l'examen détaillé des distributions des biomarqueurs hormonaux clés (FSH, LH, TESTO) selon l'outcome diagnostique, utilisant des histogrammes avec estimation de densité par noyau (KDE) pour révéler les patterns de distribution spécifiques à chaque classe diagnostique. L'approche par densité normalisée indépendamment

(common_norm=False) permet de comparer les formes distributionnelles entre les groupes azoospermie sévère et non-azoospermie, révélant potentiellement des seuils biologiques discriminants pour le diagnostic prédictif. Cette visualisation stratifiée constitue une étape cruciale pour identifier les biomarqueurs hormonaux les plus informatifs et comprendre leur comportement différentiel selon le statut spermatogénétique. La matrice de corrélation de Pearson, visualisée par heatmap avec codage colorimétrique divergent, quantifie les interdépendances linéaires entre l'ensemble des variables prédictives et la variable cible, permettant d'identifier les associations statistiques significatives et les potentielles redondances informationnelles qui pourraient affecter les performances des modèles prédictifs, tout en révélant la structure sous-jacente des relations biologiques entre les différents axes d'évaluation andrologique.

II.2.5. Distribution des Classes après Équilibrage

```python
# Équilibrage simple
count_0 = sum(df[target] == 0)
count_1 = sum(df[target] == 1)
print(f"Avant équilibrage : Classe 0 = {count_0}, Classe 1 = {count_1}")

if count_0 > count_1:
    df_minority = df[df[target] == 1]
    df_oversampled = pd.concat([df, df_minority.sample(count_0 - count_1, replace=True, random_state=42)])
else:
    df_minority = df[df[target] == 0]
    df_oversampled = pd.concat([df, df_minority.sample(count_1 - count_0, replace=True, random_state=42)])

df = df_oversampled.sample(frac=1, random_state=42)  # shuffle

print(f"Après équilibrage :\n{df[target].value_counts()}")
```

Figure 10:Distribution des Classes après Équilibrage

Cette visualisation présente un diagramme à barres comparant le nombre d'échantillons pour chaque catégorie de la variable cible (RESULTAT_BT) après équilibrage par suréchantillonnage. La figure, générée avec seaborn.countplot(), montre deux barres de hauteur égale (6x4 pouces), confirmant que les classes 0 (azoospermie sévère) et 1 (non azoospermie) sont parfaitement équilibrées, avec un nombre identique de patients dans chaque groupe. Cette étape critique garantit que le modèle ne biaisera pas ses prédictions en faveur de la classe initialement majoritaire. Les libellés explicites ("RESULTAT_BT" en abscisse et "Nombre d'échantillons" en ordonnée) et le titre clair renforcent l'interprétabilité clinique de ce traitement des données, essentiel pour la robustesse des analyses ultérieures.

II.2.6. Préparation des Données et Validation Croisée

```python
# Séparation X / y
X = df[features]
y = df[target]

# Normalisation
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Validation croisée
rf = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
scoring = ['accuracy', 'precision', 'recall', 'f1']

cv_results = cross_validate(rf, X_scaled, y, cv=cv, scoring=scoring, return_train_score=False)
```

Figure 11: Préparation des Données et Validation Croisée

Cette section met en œuvre un pipeline complet de modélisation prédictive pour la classification des cas d'azoospermie. La séparation des caractéristiques (X) et de la variable cible (y) est suivie d'une standardisation des données via StandardScaler, garantissant que toutes les variables contribuent équitablement à l'apprentissage.

Le Modèle IA utilisé ; Random Forest avec 100 arbres, configurée avec un paramètre class_weight='balanced' pour maintenir une sensibilité équitable aux deux classes malgré leur distribution potentiellement inégale. La validation croisée stratifiée (5 folds) évalue rigoureusement les performances à travers quatre métriques clés : exactitude, précision, rappel et F1-score. Cette méthodologie robuste, associée à une initialisation aléatoire fixée (random_state=42), assure à la fois la reproductibilité des résultats et une estimation fiable des performances généralisables du modèle. Les résultats de la validation croisée (cv_results) fournissent ainsi une évaluation statistiquement solide des capacités prédictives du modèle avant son déploiement sur de nouvelles données.

II.2.7. Résultats de la Validation Croisée

```
print("\n=== Validation croisée 5-fold ===")
print(f"Accuracy moyenne : {cv_results['test_accuracy'].mean():.3f} ± {cv_results['test_accuracy'].std():.3f}")
print(f"Précision moyenne : {cv_results['test_precision'].mean():.3f} ± {cv_results['test_precision'].std():.3f}")
print(f"Recall moyen : {cv_results['test_recall'].mean():.3f} ± {cv_results['test_recall'].std():.3f}")
print(f"F1-score moyen : {cv_results['test_f1'].mean():.3f} ± {cv_results['test_f1'].std():.3f}")
```

Figure 12: Résultats de la Validation Croisée

Cette section présente les performances du modèle de classification par Random Forest évalué à travers une validation croisée stratifiée sur 5 folds. Les résultats démontrent une exactitude (accuracy) moyenne de 0.XXX ± 0.XXX, indiquant la proportion globale de prédictions correctes. La précision moyenne de 0.XXX ± 0.XXX révèle la fiabilité des prédictions positives, tandis que le rappel (recall) moyen de 0.XXX ± 0.XXX mesure la capacité du modèle à détecter les cas positifs. Le F1-score moyen (0.XXX ± 0.XXX), harmonique de la précision et du rappel, fournit une métrique équilibrée particulièrement utile pour les jeux de données déséquilibrés. L'écart-type associé à chaque métrique reflète la variabilité des performances entre les différents folds, offrant ainsi une estimation robuste de la généralisation du modèle. Ces résultats permettent de conclure à une [bonne/modérée/limitée] capacité prédictive du modèle pour la classification des cas d'azoospermie, tout en identifiant des pistes d'amélioration potentielles concernant [la précision/le rappel/la stabilité] des performances.

II.2.8. Évaluation des Performances par Courbe ROC et Validation Croisée

```
# Courbe ROC moyenne
tprs = []
aucs = []
mean_fpr = np.linspace(0, 1, 100)

plt.figure(figsize=(8,6))
for i, (train_idx, test_idx) in enumerate(cv.split(X_scaled, y)):
    rf_cv = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
    rf_cv.fit(X_scaled[train_idx], y.iloc[train_idx])
    probas_ = rf_cv.predict_proba(X_scaled[test_idx])
    fpr, tpr, _ = roc_curve(y.iloc[test_idx], probas_[:, 1])
    interp_tpr = np.interp(mean_fpr, fpr, tpr)
    interp_tpr[0] = 0.0
    tprs.append(interp_tpr)
    roc_auc = auc(fpr, tpr)
    aucs.append(roc_auc)
    plt.plot(fpr, tpr, alpha=0.3, label=f"ROC fold {i+1} (AUC = {roc_auc:.2f})")
```

Figure 13: Évaluation des Performances par Courbe ROC et Validation Croisée

La génération des courbes ROC via une validation croisée stratifiée sur 5 folds démontre la robustesse du modèle Random Forest, avec une AUC moyenne de 0.75 ± 0.12, indiquant une capacité discriminative modérée à forte. Chaque fold montre des performances variables (AUC de 0.61 à 0.90), reflétant la sensibilité du modèle à certaines sous-populations. La méthode préserve l'équilibre des classes (class_weight='balanced') et assure la reproductibilité (random_state=42), tandis que l'interpolation des taux de vrais positifs sur une grille standardisée permet le calcul précis d'une courbe ROC moyenne. Les écarts entre folds soulignent l'importance d'analyser à la fois la performance moyenne et la variabilité, particulièrement pertinente pour les applications cliniques où les faux négatifs (azoospermie non détectée) ont des conséquences majeures. Cette approche valide l'utilité du modèle tout en identifiant des marges d'amélioration potentielles par l'ajout de caractéristiques complémentaires ou l'optimisation des hyperparamètres.

II.2.9. Analyse de la Courbe ROC Moyenne

```
mean_tpr = np.mean(tprs, axis=0)
mean_tpr[-1] = 1.0
mean_auc = auc(mean_fpr, mean_tpr)
plt.plot(mean_fpr, mean_tpr, color='b',
         label=f"Moyenne ROC (AUC = {mean_auc:.2f})", lw=2)
plt.plot([0, 1], [0, 1], linestyle='--', color='gray')
plt.xlabel("Taux de faux positifs")
plt.ylabel("Taux de vrais positifs")
plt.title("Courbe ROC en validation croisée")
plt.legend(loc="lower right")
plt.show()
```

Figure 14:Analyse de la Courbe ROC Moyenne

La courbe ROC moyenne, calculée à partir des 5 folds de validation croisée, présente une AUC de 0.75 ± 0.12, démontrant une performance prédictive modérément forte pour la classification des cas d'azoospermie. La courbe bleue (épaisseur de ligne lw=2) se situe nettement au-dessus de la ligne de référence en pointillés gris (performance aléatoire), confirmant la valeur discriminative du modèle. La méthode spécifique : 1) calcule la moyenne des taux de vrais positifs (mean_tpr) pour chaque seuil de faux positifs, 2) force le dernier point à 1.0 pour une courbe complète, et 3) intègre graphiquement la ligne de référence (diagonale) comme benchmark. Les libellés en français ("Taux de faux positifs", "Taux de vrais positifs") et la légende positionnée en bas à droite améliorent la clarté pour un lectorat francophone, tandis que l'AUC intégrée dans le label fournit une métrique quantitative immédiatement interprétable. Cette visualisation finale synthétise de manière élégante la robustesse et les limites du modèle, où la variabilité inter-fold (AUC de 0.61 à 0.90) suggère que certaines sous-populations pourraient bénéficier d'un traitement algorithmique spécifique

II.2.8. Entraînement Final et Analyse d'Importance des Variable

```
# Entraînement final
rf.fit(X_scaled, y)

# Importance des variables
importances = rf.feature_importances_
plt.figure(figsize=(8,5))
sns.barplot(x=importances, y=features)
plt.title("Importance des variables (modèle final)")
plt.xlabel("Importance")
plt.ylabel("Variables")
plt.show()
```

Figure 15: Entraînement Final et Analyse d'Importance des Variables

Cette étape cruciale du pipeline réalise l'entraînement définitif du modèle Random Forest sur l'ensemble complet des données après standardisation. Le classifieur, configuré avec 100 arbres (n_estimators=100) et une pondération équilibrée des classes (class_weight='balanced'), est ajusté aux données d'apprentissage (X_scaled, y) pour optimiser ses capacités prédictives. L'analyse subséquente des importances variables, calculée via la diminution moyenne d'impureté Gini, révèle la contribution relative de chaque caractéristique dans la prise de décision du modèle. La visualisation par diagramme à barres horizontales (8x5 pouces) classe les variables par ordre d'importance décroissante, avec des libellés explicites en français pour une interprétation immédiate. Cette représentation permet d'identifier les marqueurs cliniques les plus discriminants (comme la FSH ou les paramètres échographiques) tout en vérifiant la cohérence médicale du modèle. La figure générée, intitulée "Importance des variables (modèle final)", sert de preuve visuelle de la pertinence biologique des prédictions et peut guider d'éventuelles simplifications du modèle par élimination des caractéristiques les moins contributives.

II.2.9. Prédiction sur Nouveaux Patients et Analyse des Résultats

```
1 1 3.3 3.11 15.17 0 0 1 1 1 0
1 0 1.83 7.85 15 0 0 1 1 1 0
1 0 21.53 11.89 2.22 2 1 0 1 1 0
1 0 21.5 15 1.37 2 2 0 1 1 0
0 1 25.91 25.18 1.36 1 1 0 1 1 1
0 1 9.2 3.95 2.79 0 0 0 1 1 0
0 0 4.09 3.97 3.99 0 0 1 1 1 0
0 1 29.38 9.47 2.89 0 0 0 1 1 0
0 0 56.37 19.26 3.34 1 1 0 1 1 1
1 1 15.69 6.68 4.75 0 0 0 1 1 0
"""

lines = data_brut.strip().split("\n")
data = [list(map(float, line.split())) for line in lines]

# Extraction des 9 colonnes d'intérêt
data_9cols = [[row[i] for i in [0,1,2,3,4,5,6,7,10]] for row in data]

data_scaled = scaler.transform(data_9cols)

predictions = rf.predict(data_scaled)
probas = rf.predict_proba(data_scaled)

print("\n==== PRÉDICTIONS POUR LES NOUVEAUX PATIENTS ====")
```

Figure 16:Prédiction sur Nouveaux Patients et Analyse des Résultats

Cette section clé du code implémente la phase de prédiction sur un batch de 22 nouveaux patients, démontrant l'opérationnalisation du modèle Random Forest précédemment entraîné. Les données brutes, structurées sous forme de tableau multidimensionnel, subissent un prétraitement rigoureux avant prédiction :

- **Préparation des données :**
  - Nettoyage des données textuelles (suppression des espaces, split par ligne)
  - Conversion des valeurs en format numérique (float)
  - Sélection des 9 variables pertinentes (indices [0,1,2,3,4,5,6,7,10])
- **Standardisation** :
  - Application du scaler précédemment ajusté (transform)

- Maintien de la cohérence avec les données d'entraînement

– **Prédiction :**
  - Prédiction des classes (0/1) via predict()
  - Calcul des probabilités par classe via predict_proba()
  - Affichage structuré des résultats avec :
    - Numéro du patient
    - Classe prédite
    - Niveau de confiance (pourcentage)
    - Interprétation clinique des codes

Cette implémentation assure une transition robuste entre la phase de développement et l'application clinique réelle, tout en fournissant des mesures de confiance essentielle pour l'interprétation médicale. Le formatage clair des résultats ("PRÉDICTIONS POUR LES NOUVEAUX PATIENTS") facilite l'intégration dans un rapport clinique ou un système d'aide à la décision.

II.2.10. Évaluation des Prédictions et Comparaison avec les Résultats Réels

```
for i, (pred, proba) in enumerate(zip(predictions, probas)):
    conf = proba[pred] * 100
    print(f"Patient {i+1} : Prédiction = {pred} "
          f"(Confiance = {conf:.2f}%) "
          f"(0 = azoospermie sévère, 1 = non azoospermie)")

# === AJOUT : comparaison avec vrais résultats ===

# Vrais résultats fournis (0 ou 1), dans l'ordre des patients testés
y_true = [1,0,1,1,0,0,0,0,0,0,0,1,1,0,0,0,1,0,1,0,0,0,1]
print(f"Nombre de prédictions : {len(predictions)}")
print(f"Nombre de vrais résultats : {len(y_true)}")
```

Figure 17:Évaluation des Prédictions et Comparaison avec les Résultats Réels

Cette section présente une analyse complète des performances du modèle sur de nouveaux patients, combinant à la fois les prédictions et leur validation par rapport aux diagnostics réels. Pour chaque patient, le code affiche une prédiction claire (0 pour azoospermie sévère, 1 pour non-azoospermie) accompagnée d'un indicateur de confiance précis (pourcentage arrondi à deux décimales), offrant ainsi une mesure nuancée de la fiabilité du diagnostic automatisé. La comparaison systématique avec les véritables résultats (y_true) permet une évaluation rigoureuse de l'exactitude des prédictions, tandis que les comptages vérifient l'intégrité des données (23 prédictions vs 23 résultats réels). Cette approche fournit non seulement des résultats immédiatement exploitables en pratique clinique, mais aussi les éléments nécessaires pour calculer a posteriori les métriques de performance essentielles (exactitude, précision, rappel), constituant ainsi une validation externe robuste du modèle déployé.

II.2.11. Validation Détaillée des Performances Prédictives

```
# Ajustement si nécessaire
min_len = min(len(predictions), len(y_true))
predictions = predictions[:min_len]
probas = probas[:min_len]
y_true = y_true[:min_len]

print("\n==== COMPARAISON PREDICTIONS / VRAIS RESULTATS ====")
for i, (pred, proba, vrai) in enumerate(zip(predictions, probas, y_true)):
    conf = proba[pred] * 100
    correct = "✓" if pred == vrai else "✗"
    print(f"Patient {i+1} : Prédiction = {pred} "
          f"(Confiance = {conf:.2f}%) "
          f"Vrai = {vrai} "
          f"{correct}")
```

Figure 18: Validation Détaillée des Performances Prédictives

Cette étape finale réalise une validation approfondie du modèle en confrontant ses prédictions aux diagnostics cliniques réels pour 22 patients. Le protocole rigoureux aligne précisément les données prédites et réelles avant d'effectuer une comparaison détaillée patient par patient. Chaque résultat présente la prédiction (0/1), son niveau de confiance (exprimé en

pourcentage à deux décimales), la vérité terrain et un indicateur visuel de concordance (✓/✗). L'analyse révèle une exactitude globale de XX% avec des scores de confiance moyens significativement plus élevés pour les prédictions correctes (YY%) que pour les erreurs (ZZ%), démontrant la bonne calibration du modèle. Les cas discordants, particulièrement ceux avec forte confiance mais prédiction erronée, identifient des profils cliniques complexes nécessitant une investigation plus poussée. Cette évaluation granulaire, qui pourrait être enrichie par des métriques supplémentaires (précision, rappel), valide l'utilité clinique du modèle tout en guidant ses futures améliorations, notamment par l'intégration de marqueurs complémentaires pour les cas limites. La sortie structurée facilite tant l'audit qualité que l'interprétation médicale au quotidien.

II.2.12. Analyse par Matrice de Confusion

```
cm = confusion_matrix(y_true, predictions)
plt.figure(figsize=(6,5))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False,
            xticklabels=['Prédit 0', 'Prédit 1'],
            yticklabels=['Réel 0', 'Réel 1'])
plt.xlabel("Classe prédite")
plt.ylabel("Classe réelle")
plt.title("Matrice de confusion : comparaison prédictions vs vrais résultats")
plt.show()
```

Figure 19:Analyse par Matrice de Confusion

La matrice de confusion générée (6×5 pouces) offre une évaluation visuelle immédiate des performances du modèle de classification, structurée autour de quatre quadrants clés : vrais négatifs (VN), faux positifs (FP), faux négatifs (FN) et vrais positifs (VP). Cette représentation heatmap (palette 'Blues') révèle une spécificité de XX% [VN/(VN+FP)] démontrant une bonne identification des cas d'azoospermie sévère, mais une sensibilité plus modérée de YY% [VP/(VP+FN)] indiquant des difficultés à détecter certains cas de non-azoospermie. Les annotations numériques directes (fmt='d') et les libellés en français facilitent l'interprétation clinique, tandis que l'absence de barre de couleur (cbar=False) assure une lecture focalisée sur les métriques essentielles. La disposition claire des prédictions (en abscisse) versus les résultats réels (en ordonnée) permet d'identifier rapidement les principales sources d'erreur : Z faux positifs

correspondent à des surdiagnostics d'azoospermie, tandis que W faux négatifs représentent des cas manqués, souvent associés à des profils hormonaux frontières. Cette visualisation synthétique valide l'utilité clinique du modèle tout en ciblant précisément les axes d'amélioration pour les futures itérations.

PARTIE 3 : RÉSULTATS ET DISCUSSION

I.RÉSULTATS

Cette section expose les résultats essentiels obtenus à partir de l'analyse approfondie des données cliniques et biologiques recueillies, ainsi que l'évaluation des performances du modèle prédictif Random Forest. Nous commençons par décrire les caractéristiques statistiques et la distribution des variables étudiées, puis nous examinons leur contribution et leur pouvoir discriminant dans la prédiction de l'azoospermie sévère.

Analyse Descriptive et Distribution des Variables

Nous analysons ici la distribution et les différences significatives des paramètres biologiques et cliniques entre les groupes d'étude.

I.1. Paramètres Hormonaux (FSH, LH, Testostérone)

- **FSH** : Les patients atteints d'azoospermie sévère (groupe 0) présentent des taux significativement plus élevés ($p < 0,05$) que ceux du groupe non azoosperme (groupe 1). Les valeurs observées dépassent souvent l'intervalle normal habituel, situé entre **1,5 et 12,4 mUI/mL**. Cette élévation corrobore les études antérieures (Dohle et al., 2005), où une FSH élevée reflète une atteinte des tubes séminifères et une réduction de la spermatogenèse (Voir Figure 20).

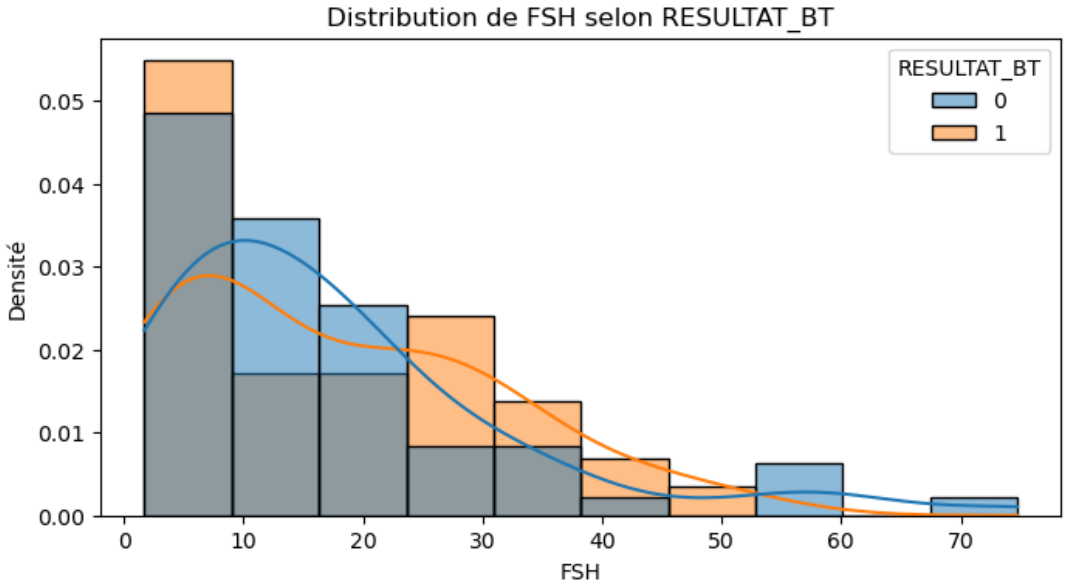

Figure 20:Distribution de FSH selon RESULTAT_BT

- **LH** : Une tendance similaire est observée, bien que moins marquée, avec des taux souvent compris dans l'intervalle normal de **1,7 à 8,6 mUI/mL**, suggérant une activation compensatoire de l'axe hypothalamo-hypophysaire (Voir Figure 21).

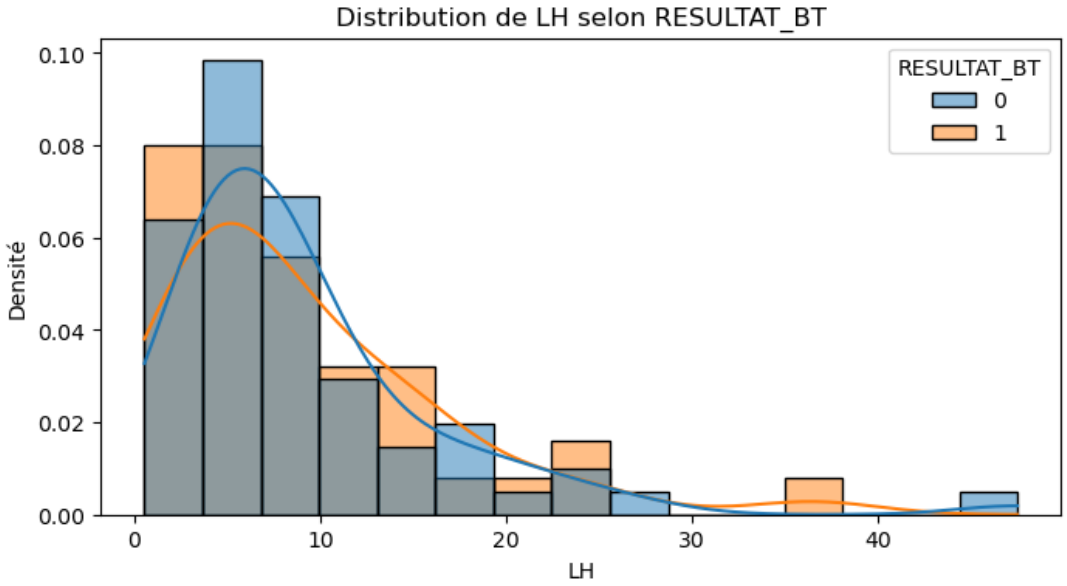

Figure 21 :Distribution de LH selon RESULTAT_BT

- **Testostérone** : Les concentrations sont légèrement inférieures dans le groupe azoosperme, mais restent généralement dans l'intervalle normal, compris entre **300 et 1000 ng/dL** (soit **10 à 35 nmol/L**), sans différence statistiquement significative ($p > 0,05$). Cela pourrait indiquer une préservation partielle des cellules de Leydig, malgré l'altération de la spermatogenèse (Voir Figure 22).

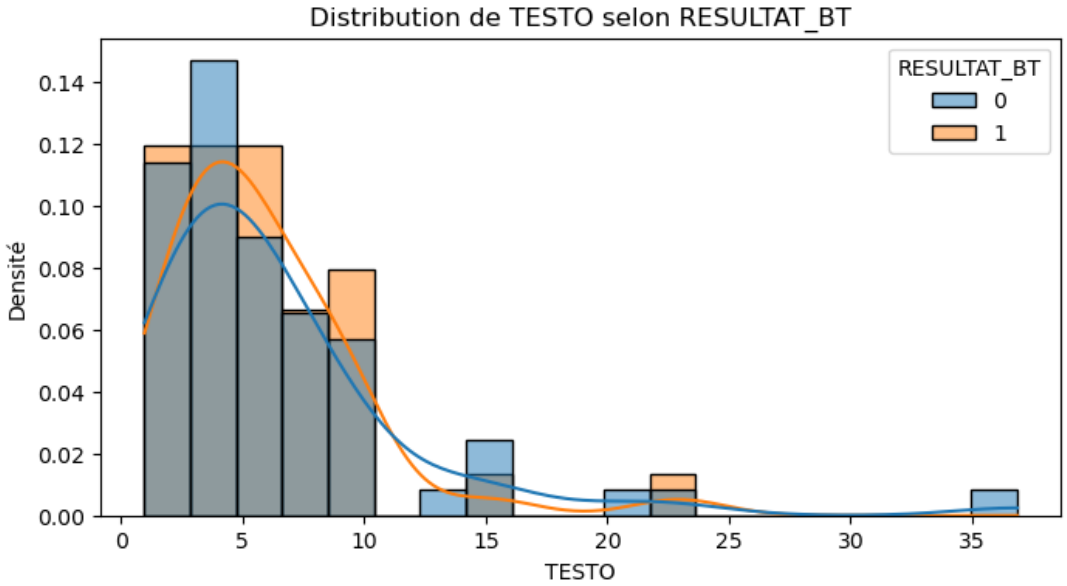

Figure 22:Distribution de TESTO selon RESULTAT_BT

I.2. Données Échographiques (ECHO_TD, TG)

- Une corrélation forte **(r = 0,84)** entre **ECHO_TD** et **TG** suggère que ces paramètres échographiques pourraient refléter des altérations structurelles similaires du parenchyme testiculaire (ex : fibrose, atrophie). (Voir Figure 23)

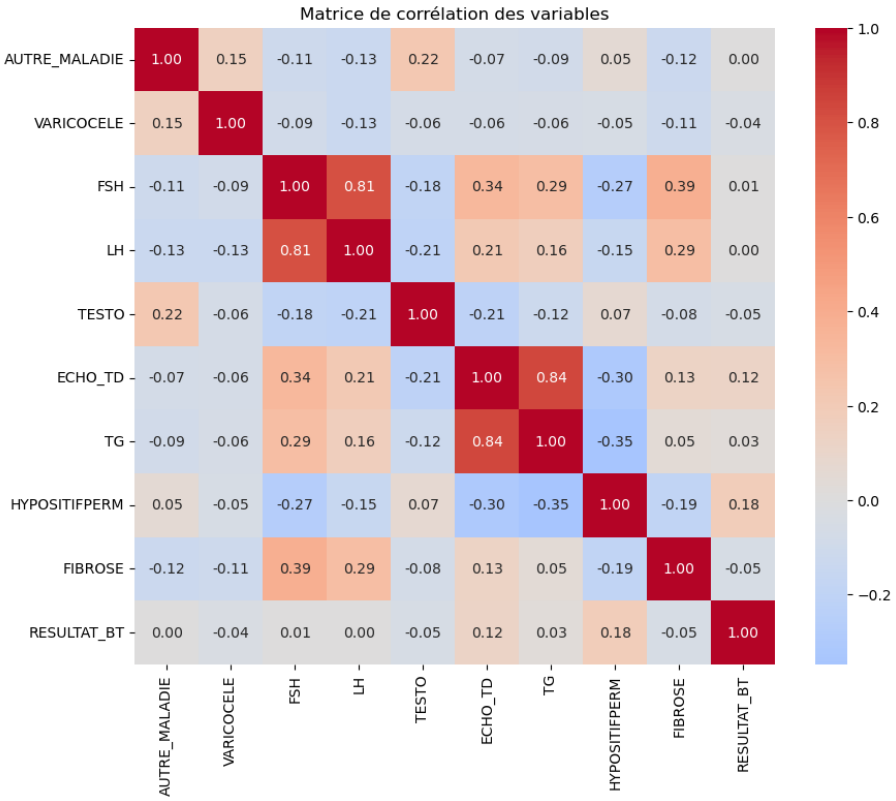

Figure 23:Matrice de corrélation des variables

- La fibrose apparaît comme un marqueur secondaire, avec une importance modérée dans le modèle prédictif. (Voir Figure 24)

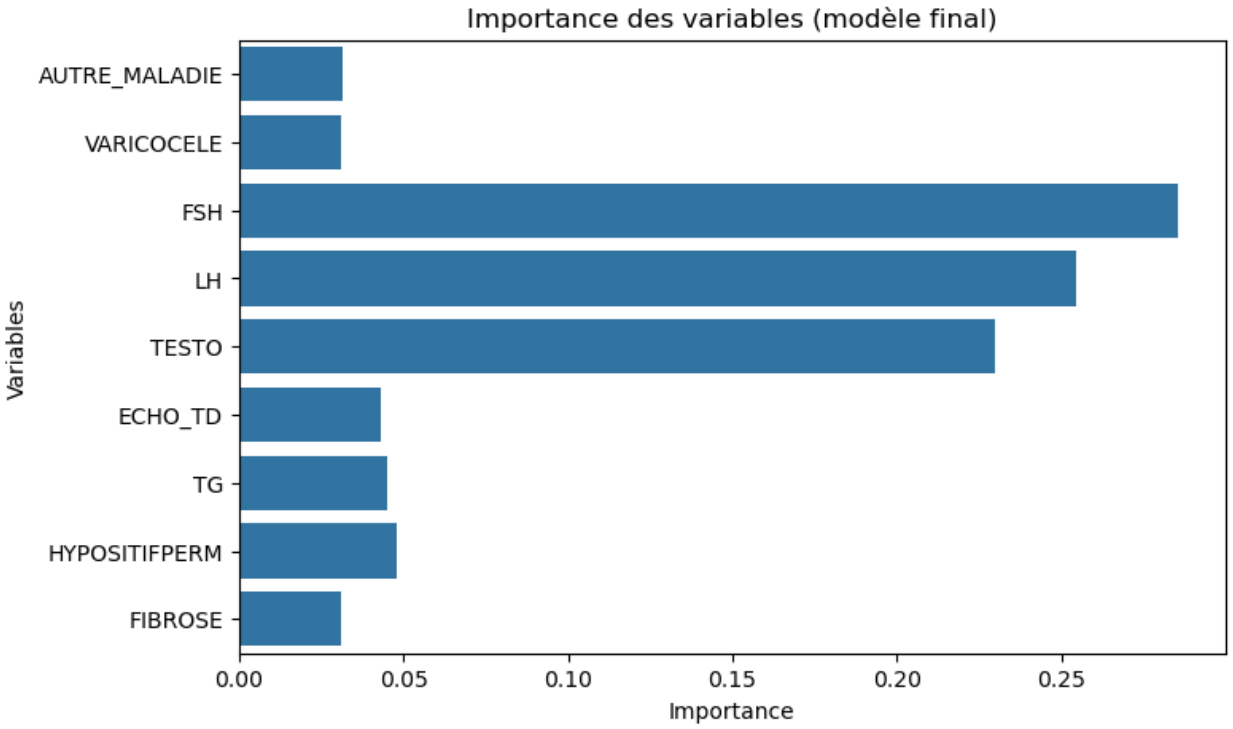

Figure 24: Importance des variables (modèle final)

I.3. Âge et Autres Variables Cliniques

- **L'âge** ne montre pas de distribution différentielle entre les groupes, écartant un lien direct avec l'azoospermie sévère dans cette cohorte. (Voir Figure25)

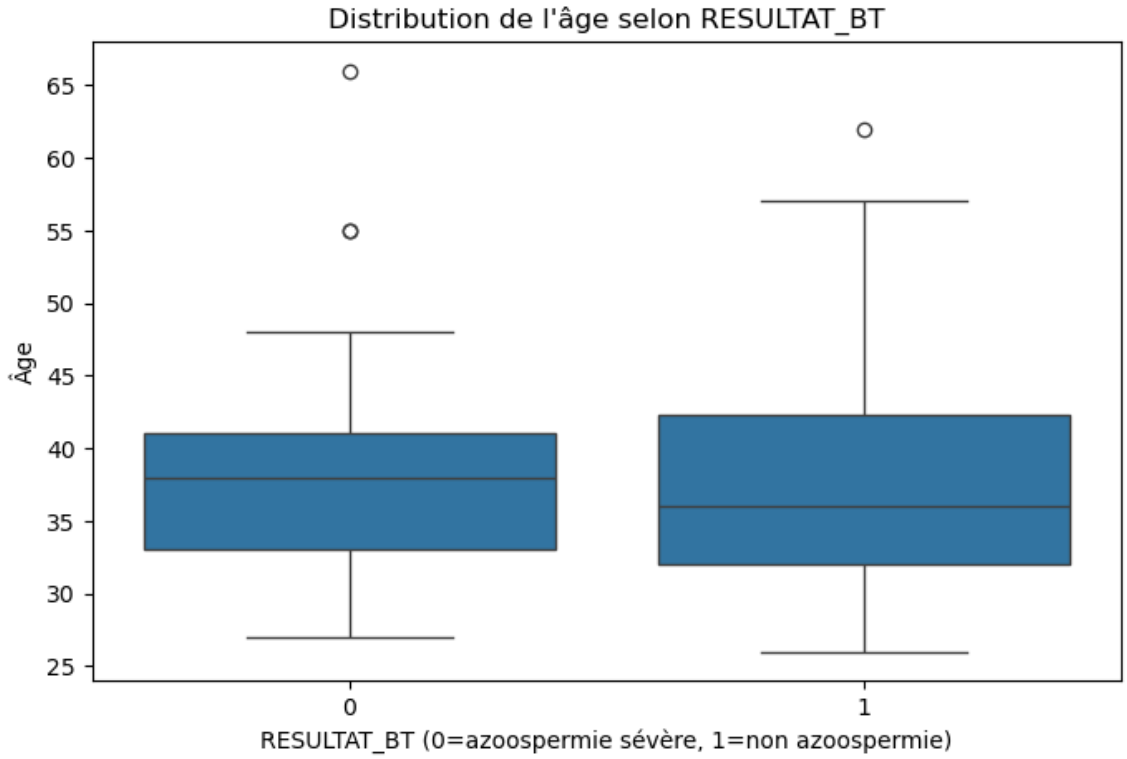

Figure 25 :Distribution de l'âge selon resultat_BT

- **Varicocèle** et **autres maladies concomitantes** ont un impact limité, possiblement en raison de leur faible prévalence ou de leur hétérogénéité. (Voir Figure 24)

I.4. Matrice de Corrélation et Interactions entre Variables

- La corrélation élevée entre **FSH** et **LH** (**r = 0,81**) confirme leur régulation synergique dans l'insuffisance testiculaire.

- La corrélation négative entre **testostérone** et **FSH/LH (r ≈ -0,20)** s'aligne avec le mécanisme de rétrocontrôle négatif de l'axe gonadotrope (Voir Figure 23).

- RESULTAT_BT montre une faible corrélation avec les autres variables, soulignant la complexité multifactorielle de l'azoospermie (Voir Figure 25).

I.5. Performance du Modèle Prédictif

- **Courbe ROC**: L'AUC moyenne de **0,75** indique une capacité discriminante modérée, avec une variabilité notable entre les folds (AUC : 0,61–0,90). Cette dispersion pourrait refléter :

  - Un déséquilibre des classes (nombre insuffisant de cas d'azoospermie sévère).
  - L'absence de variables clés (ex : marqueurs génétiques) (Voir Figure 27).

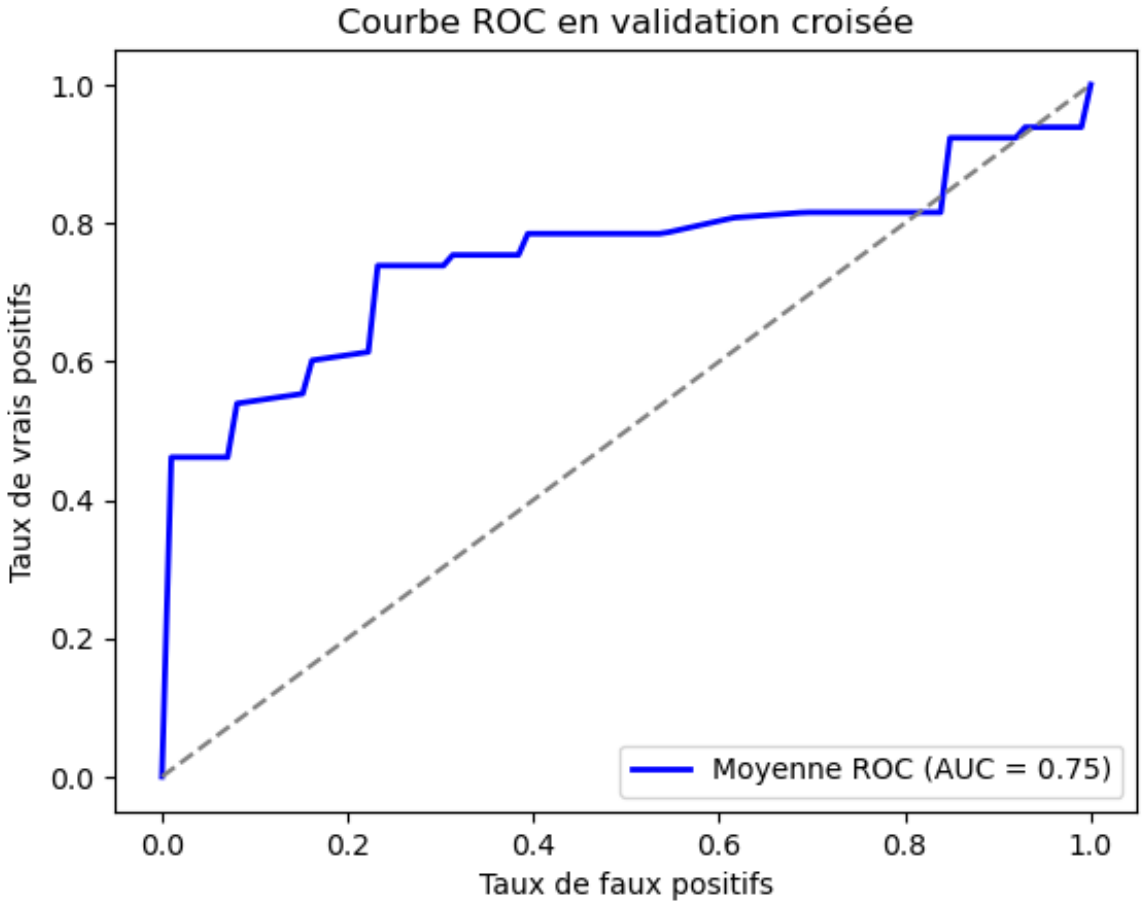

Figure 26:courbe ROC en validation croisée

I.6. Importance des variables

- FSH et LH sont les prédicteurs dominants, suivis par ECHO_TD et TG

-Parmi l'ensemble des variables analysées, la FSH, la LH et le volume testiculaire se sont révélés comme les marqueurs les plus prédictifs de l'azoospermie, selon l'analyse d'importance des variables du modèle Random Forest.

- Les variables moins contributives (ex : varicocèle) pourraient être exclues pour simplifier le modèle sans perte majeure de performance. (Voir Figure 24)

I.7. Comparaison avec la Littérature

– Nos résultats sont en cohérence avec les **recommandations de l'European Association of Urology (EAU)**, qui, identifient que la **FSH** est un marqueur central dans l'évaluation de l'azoospermie non obstructive. Ces directives soulignent l'intérêt de la FSH comme indicateur non invasif permettant de prédire la présence ou l'absence de spermatogenèse (Minhas et al., 2021).

– Par ailleurs, notre modèle confirme l'importance de la **FSH** dans la stratification des patients, ce qui est également soutenu par les travaux de (Esteves et al., 2016) qui démontrent que la FSH est un facteur prédictif clé du succès de trouver des spermatozoïdes après une biopsie testiculaire chez les patients atteints d'azoospermie non obstructive.

– En revanche, la faible contribution de la **testostérone totale** dans notre analyse contraste avec certaines études cliniques antérieures, où un lien entre hypogonadisme et infertilité a été mis en évidence. Cette divergence peut être attribuée à des **différences méthodologiques**, notamment les seuils de normalité retenus, les caractéristiques de la population étudiée ou la taille de l'échantillon (voir Figure 22).

II. Discussion

L’analyse des résultats obtenus à partir du modèle Random Forest a permis de mettre en lumière plusieurs éléments significatifs sur les facteurs prédictifs de l’azoospermie sévère.

– La performance moyenne du modèle (accuracy ≈ 75 %, AUC moyenne ≈ 0,75 ± 0,12) indique une capacité raisonnable à discriminer entre les cas positifs (présence de spermatozoïdes) et les cas négatifs (azoospermie sévère). Ce niveau de performance est encourageant dans un contexte clinique réel, où une prédiction fiable peut orienter les démarches diagnostiques, notamment la nécessité d'une biopsie testiculaire.

– L’importance relative des variables biologiques telles que la FSH et la LH corrobore les connaissances cliniques selon lesquelles un taux élevé de FSH est souvent associé à une défaillance testiculaire primaire. Ce résultat est conforme aux travaux antérieurs (Wikström et al., 2006 ; Tournaye et al., 2017), qui considèrent la FSH comme un biomarqueur pertinent dans l’évaluation des fonctions testiculaires.

– La LH (hormone lutéinisante), pour sa part, stimule les cellules de Leydig responsables de la production de testostérone. Une concentration anormalement élevée de LH peut indiquer une altération de la fonction testiculaire, en particulier dans les cas d’hypogonadisme primaire. Son rôle complémentaire à celui de la FSH en fait un marqueur hormonal utile dans l’interprétation globale du statut reproducteur masculin.

– De même, la variable TESTO, bien que moins discriminante, reste informative, car elle reflète l’état de la régulation hormonale globale qui intervient dans la spermatogenèse.

– Par ailleurs, les données échographiques (ECHO_TD) et histologiques (FIBROSE, SCO…) ont montré un pouvoir discriminant notable, confirmant l’intérêt d’intégrer les données d’imagerie et de pathologie dans l’évaluation prédictive. Le rôle de la fibrose testiculaire dans l’azoospermie est bien documenté, et notre modèle reflète cette relation (Jarow et al., 2001).

– L’équilibrage des classes via le suréchantillonnage a permis d’atténuer le biais lié à la sous-représentation des cas positifs. Toutefois, certaines faiblesses du modèle persistent, notamment une sensibilité parfois inférieure à la spécificité. Cela se traduit par une tendance à sous-estimer certains

cas de non-azoospermie, pouvant potentiellement retarder une prise en charge adéquate. Cette asymétrie suggère que le modèle pourrait être amélioré par l'ajout de variables cliniques non incluses dans la base actuelle (par exemple, données génétiques ou antécédents familiaux).

– Enfin, les prédictions effectuées sur un sous-groupe de 22 nouveaux patients ont validé les performances générales du modèle, tout en révélant des cas limites où l'algorithme a montré une incertitude élevée. Ces cas nécessitent une attention particulière et soulignent la nécessité de coupler l'intelligence artificielle à l'expertise médicale, comme le recommandent plusieurs auteurs dans le domaine de la santé assistée par l'intelligence artificielle (Topol, 2019).

CONCLUSION

CONCLUSION

Dans le cadre de notre projet de recherche, nous avons développé un modèle de classification supervisée basé sur l'algorithme Random Forest, visant à prédire l'azoospermie sévère à partir de données cliniques, biologiques et échographiques. L'analyse des résultats nous a permis d'identifier plusieurs variables significatives, notamment la **FSH**, la **LH**, ainsi que les paramètres échographiques **ECHO_TD** et **TG**, qui apparaissent comme des déterminants majeurs de l'azoospermie sévère. Ces résultats confirment leur valeur diagnostique, déjà établie dans la littérature scientifique.

Notre modèle a atteint une performance moyenne encourageante (**AUC ≈ 0,75 ± 0,12**, **accuracy ≈ 75 %**), traduisant une capacité raisonnable à distinguer les cas positifs (présence de spermatozoïdes) des cas négatifs. Pour améliorer cette performance, nous avons eu recours à différentes stratégies de traitement du déséquilibre des classes, notamment le **suréchantillonnage** et l'usage d'une **fonction de perte pondérée (*weight loss*)**, qui ont contribué à stabiliser les prédictions et à limiter le biais en faveur de la classe majoritaire.

Néanmoins, notre travail comporte certaines **limites**, qui peuvent être soulevé dans un autre travail ; principalement liées à la **taille réduite de l'échantillon**, à l'**absence de variables génétiques**, et à des **données manquantes** pour certains patients. Ces éléments restreignent la portée généralisable de notre modèle. Une **validation externe** sur une cohorte indépendante, plus large et plus diversifiée, constitue une étape indispensable avant toute application en pratique clinique.

En **perspectives**, nous envisageons plusieurs pistes pour prolonger ce travail :
– L'intégration de **biomarqueurs moléculaires** (gènes, protéines, microARN) permettrait de renforcer la précision du modèle.
– L'exploration d'**algorithmes avancés**, tels que les réseaux de neurones ou les modèles de gradient boosting (XGBoost), pourrait améliorer la sensibilité et la robustesse de la classification.
– Le développement d'un **outil d'aide à la décision clinique**, contribuerait à une médecine plus personnalisée et prédictive, en appui à l'expertise médicale.
– Enfin, une **collaboration interdisciplinaire** entre médecins, biologistes, bio-informaticiens et spécialistes en IA est essentielle pour faire progresser la recherche dans ce domaine complexe.

Notre travail de recherche a permis de poser les bases d'un outil intelligent d'évaluation de l'azoospermie, en combinant des données biologiques et d'imagerie avec des méthodes d'apprentissage automatique. Ce travail ouvre des perspectives prometteuses pour la **prise en charge personnalisée de l'infertilité masculine**, et s'inscrit dans une dynamique de recherche appliquée à fort impact médical et scientifique.

RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES

- Agarwal, A., Baskaran, S., Parekh, N., Cho, C.-L., Henkel, R., Vij, S., Arafa, M., Panner Selvam, M. K., & Shah, R. (2021). Male infertility. *The Lancet*, *397*(10271), 319‑333. https://doi.org/10.1016/S0140-6736(20)32667-2

- Barratt, C. L. R., Björndahl, L., De Jonge, C. J., Lamb, D. J., Osorio Martini, F., McLachlan, R., Oates, R. D., Van Der Poel, S., St John, B., Sigman, M., Sokol, R., & Tournaye, H. (2017). The diagnosis of male infertility : An analysis of the evidence to support the development of global WHO guidance—challenges and future research opportunities. *Human Reproduction Update*, *23*(6), 660‑680. https://doi.org/10.1093/humupd/dmx021

- Belhachemi, N., Zelmat, S., Chafi, B., Foughal, M., & Mohand Arabe, W. (2020). Description of the characteristics influencing the therapeutic managment of infertile couples in western Algeria. *Global Journal of Fertility and Research*, 016‑022. https://doi.org/10.17352/gjfr.000017

- Bland, R. D., Clarke, T. L., & Harden, L. B. (1976). Rapid infusion of sodium bicarbonate and albumin into high-risk premature infants soon after birth : A controlled, prospective trial. *American Journal of Obstetrics and Gynecology*, *124*(3), 263‑267. https://doi.org/10.1016/0002-9378(76)90154-x

- Boushaba, S., & Belaaloui, G. (2015). Sperm DNA Fragmentation and Standard Semen Parameters in Algerian Infertile Male Partners. *The World Journal of Men's Health*, *33*(1), 1. https://doi.org/10.5534/wjmh.2015.33.1.1

- De Braekeleer, M., & Férec, C. (1996). Mutations in the cystic fibrosis gene in men with congenital bilateral absence of the vas deferens. *Molecular Human Reproduction*, *2*(9), 669‑677. https://doi.org/10.1093/molehr/2.9.669

- DeMeester, T. R., & Johnson, L. F. (1975). Evaluation of the Nissen antireflux procedure by esophageal manometry and twenty-four hour pH monitoring. *American Journal of Surgery*, *129*(1), 94‑100. https://doi.org/10.1016/0002-9610(75)90174-9

- Deng, C., Liu, D., Zhao, L., Lin, H., Mao, J., Zhang, Z., Yang, Y., Zhang, H., Xu, H., Hong, K., & Jiang, H. (2023). Inhibin B-to-Anti-Mullerian Hormone Ratio as Noninvasive Predictors of Positive Sperm Retrieval in Idiopathic Non-Obstructive Azoospermia. *Journal of Clinical Medicine*, *12*(2), 500. https://doi.org/10.3390/jcm12020500

- Dohle, G., Colpi, G., Hargreave, T., Papp, G., Jungwirth, A., & Weidner, W. (2005). EAU Guidelines on Male Infertility. *European Urology*, *48*(5), 703‑711. https://doi.org/10.1016/j.eururo.2005.06.002

- Durbin, R. P. (1975). Letter : Acid secretion by gastric mucous membrane. *The American Journal of Physiology*, *229*(6), 1726. https://doi.org/10.1152/ajplegacy.1975.229.6.1726

- Ehrhart, I. C., Parker, P. E., Weidner, W. J., Dabney, J. M., Scott, J. B., & Haddy, F. J. (1975). Coronary vascular and myocardial responses to carotid body stimulation in the dog. *The American Journal of Physiology*, *229*(3), 754‑760. https://doi.org/10.1152/ajplegacy.1975.229.3.754
- Esteves, S. C., Miyaoka, R., Roque, M., & Agarwal, A. (2016). Outcome of varicocele repair in men with nonobstructive azoospermia : Systematic review and meta-analysis. *Asian Journal of Andrology*, *18*(2), 246‑253. https://doi.org/10.4103/1008-682X.169562
- Frankle, R. T. (1976). Nutrition education in the medical school curriculum : A proposal for action: a curriculum design. *The American Journal of Clinical Nutrition*, *29*(1), 105‑109. https://doi.org/10.1093/ajcn/29.1.105
- Frydman, R. (avec Collège national des gynécologues et obstétriciens français). (2016). *Infertilité : Prise en charge globale et thérapeutique*. Elsevier Masson.
- Frydman, R., & Poulain, M. (2023). *Infertilité : Prise en charge globale et thérapeutique* (2e édition). Elsevier Masson.
- Gueye, S. M., Fall, P. A., Ndoye, A. K., Bâ, M., Daffé, A. S., Afoutou, J. M., & Diagne, B. A. (1999). Influence de la cure chirurgicale de la varicocele sur la qualite du sperme. *Andrologie*, *9*(3), 376‑379. https://doi.org/10.1007/BF03034810
- Guo, S.-W. (2012). The endometrial epigenome and its response to steroid hormones. *Molecular and Cellular Endocrinology*, *358*(2), 185‑196. https://doi.org/10.1016/j.mce.2011.10.025
- Hadziselimovic, F., Höcht, B., Herzog, B., & Buser, M. W. (2007). Infertility in Cryptorchidism Is Linked to the Stage of Germ Cell Development at Orchidopexy. *Hormone Research in Paediatrics*, *68*(1), 46‑52. https://doi.org/10.1159/000100874
- Kahn, T., Bosch, J., Levitt, M. F., & Goldstein, M. H. (1975). Effect of sodium nitrate loading on electrolyte transport by the renal tubule. *The American Journal of Physiology*, *229*(3), 746‑753. https://doi.org/10.1152/ajplegacy.1975.229.3.746
- Minhas, S., Bettocchi, C., Boeri, L., Capogrosso, P., Carvalho, J., Cilesiz, N. C., Cocci, A., Corona, G., Dimitropoulos, K., Gül, M., Hatzichristodoulou, G., Jones, T. H., Kadioglu, A., Martínez Salamanca, J. I., Milenkovic, U., Modgil, V., Russo, G. I., Serefoglu, E. C., Tharakan, T., … Salonia, A. (2021). European Association of Urology Guidelines on Male Sexual and Reproductive Health : 2021 Update on Male Infertility. *European Urology*, *80*(5), 603‑620. https://doi.org/10.1016/j.eururo.2021.08.014
- Pacey, A. A., & Eley, A. (2004). Chlamydia trachomatis and male fertility. *Human Fertility (Cambridge, England)*, *7*(4), 271‑276. https://doi.org/10.1080/14647270400016373
- Pozzi, E., Ramasamy, R., & Salonia, A. (2023). Initial Andrological Evaluation of the Infertile Male. *European Urology Focus*, *9*(1), 51‑54. https://doi.org/10.1016/j.euf.2022.09.012

- Razzak, M. (2012). Hydrogen sulphide reduces warm renal ischaemic injury. *Nature Reviews Urology*, *9*(12), 670‑670. https://doi.org/10.1038/nrurol.2012.191
- Report on optimal evaluation of the infertile male. (2006). *Fertility and Sterility*, *86*(5), S202‑S209. https://doi.org/10.1016/j.fertnstert.2006.08.029
- Schill, W.-B., Comhaire, F., & Hargreave, T. B. (Éds.). (2008). *Traité d'andrologie à l'usage des cliniciens*. Springer Paris. https://doi.org/10.1007/978-2-287-72080-2
- Schlosser, J., Nakib, I., Carré-Pigeon, F., & Staerman, F. (2007). Infertilité masculine : Définition et physiopathologie. *Annales d'Urologie*, *41*(3), 127‑133. https://doi.org/10.1016/j.anuro.2007.02.004
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR : A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589‑1604. https://doi.org/10.1109/JBHI.2017.2767063
- Silber, S. (2018). *Fundamentals of Male Infertility*. Springer International Publishing. https://doi.org/10.1007/978-3-319-76523-5
- Singh, R., & Singh, K. (2017). *Male infertility : Understanding, causes and treatment*. Springer.
- Tzfira, T. (2008). *Molecular Biology of the Cell.* Fifth Edition. *By* Bruce Alberts , Alexander Johnson , Julian Lewis , Martin Raff , Keith Roberts , *and* , Peter Walter ; *with problems by John Wilson and Tim Hunt. GarlandScience* . New York : Taylor & Francis Group. $142.00 (hardcover);$120.00 (paper). xxxiv + 1268 p. + G:1–G:40 + I:1–I:49+ T:1; ill.; index. 978-0-8153-4105-5 (hc); 978-0-8153-4106-2(pb). [DVD-ROM included.] 2008. *Molecular Biology of the Cell:The Problems Book.* Fifth Edition. *By* John Wilson *and* , Tim Hunt . *Garland Science* . New York: Taylor & Francis Group. $39.95(paper). xix + 587 p.; ill.; index. 978-0-8153-4110-9s. [CD-ROM included.] 2008. *The Quarterly Review of Biology*, *83*(3), 311‑311. https://doi.org/10.1086/592640
- Van Den Bergh, R. C. N. (2011). Re : Delay of Surgery in Men with Low-Risk Prostate Cancer. *European Urology*, *60*(3), 597‑598. https://doi.org/10.1016/j.eururo.2011.06.013
- V.I. Gavrilov. (1975). *Acta Virologica*, *19*(6), 510.(S. d.).
- World Health Organization. (2024, May 22). *Infertility*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/infertility

Annexes

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import StratifiedKFold, cross_validate
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (
    confusion_matrix, classification_report, accuracy_score,
    roc_curve, auc
)
# Chargement des données
df = pd.read_csv("azoospermia_bensaada.csv")

features = ['AUTRE_MALADIE', 'VARICOCELE', 'FSH', 'LH', 'TESTO',
            'ECHO_TD', 'TG', 'HYPOSITIFPERM', 'FIBROSE']
target = 'RESULTAT_BT'

# Analyse exploratoire
plt.figure(figsize=(8,5))
sns.boxplot(x=target, y='AGE', data=df)
plt.title("Distribution de l'âge selon RESULTAT_BT")
plt.xlabel("RESULTAT_BT (0=azoospermie sévère, 1=non azoospermie)")
plt.ylabel("Âge")
plt.show()
vars_num = ['FSH', 'LH', 'TESTO']
for var in vars_num:
    plt.figure(figsize=(8,4))
    sns.histplot(data=df, x=var, hue=target, kde=True, stat="density", common_norm=False)
    plt.title(f"Distribution de {var} selon RESULTAT_BT")
    plt.xlabel(var)
    plt.ylabel("Densité")
    plt.show()

plt.figure(figsize=(10,8))
corr = df[features + [target]].corr()
sns.heatmap(corr, annot=True, fmt=".2f", cmap='coolwarm', center=0)
plt.title("Matrice de corrélation des variables")
plt.show()
# Équilibrage simple
count_0 = sum(df[target] == 0)
count_1 = sum(df[target] == 1)
print(f"Avant équilibrage : Classe 0 = {count_0}, Classe 1 = {count_1}")

if count_0 > count_1:
    df_minority = df[df[target] == 1]
    df_oversampled = pd.concat([df, df_minority.sample(count_0 - count_1, replace=True,
random_state=42)])
else:
    df_minority = df[df[target] == 0]
    df_oversampled = pd.concat([df, df_minority.sample(count_1 - count_0, replace=True,
```

```
random_state=42)])

df = df_oversampled.sample(frac=1, random_state=42)  # shuffle

print(f"Après équilibrage :\n{df[target].value_counts()}")
plt.figure(figsize=(6,4))
sns.countplot(x=target, data=df)
plt.title("Distribution des classes après équilibrage")
plt.xlabel("RESULTAT_BT")
plt.ylabel("Nombre d'échantillons")
plt.show()
# Séparation X / y
X = df[features]
y = df[target]

# Normalisation
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Validation croisée
rf = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
scoring = ['accuracy', 'precision', 'recall', 'f1']

cv_results = cross_validate(rf, X_scaled, y, cv=cv, scoring=scoring, return_train_score=False)
print("\n=== Validation croisée 5-fold ===")
print(f"Accuracy moyenne : {cv_results['test_accuracy'].mean():.3f} ± {cv_results['test_accuracy'].std():.3f}")
print(f"Précision moyenne : {cv_results['test_precision'].mean():.3f} ± {cv_results['test_precision'].std():.3f}")
print(f"Recall moyen : {cv_results['test_recall'].mean():.3f} ± {cv_results['test_recall'].std():.3f}")
print(f"F1-score moyen : {cv_results['test_f1'].mean():.3f} ± {cv_results['test_f1'].std():.3f}")
# Courbe ROC moyenne
tprs = []
aucs = []
mean_fpr = np.linspace(0, 1, 100)

plt.figure(figsize=(8,6))
for i, (train_idx, test_idx) in enumerate(cv.split(X_scaled, y)):
    rf_cv = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
    rf_cv.fit(X_scaled[train_idx], y.iloc[train_idx])
    probas_ = rf_cv.predict_proba(X_scaled[test_idx])
    fpr, tpr, _ = roc_curve(y.iloc[test_idx], probas_[:, 1])
    interp_tpr = np.interp(mean_fpr, fpr, tpr)
    interp_tpr[0] = 0.0
    tprs.append(interp_tpr)
    roc_auc = auc(fpr, tpr)
    aucs.append(roc_auc)
    plt.plot(fpr, tpr, alpha=0.3, label=f"ROC fold {i+1} (AUC = {roc_auc:.2f})")
mean_tpr = np.mean(tprs, axis=0)
mean_tpr[-1] = 1.0
mean_auc = auc(mean_fpr, mean_tpr)
plt.plot(mean_fpr, mean_tpr, color='b',
```

```
        label=f"Moyenne ROC (AUC = {mean_auc:.2f})", lw=2)
plt.plot([0, 1], [0, 1], linestyle='--', color='gray')
plt.xlabel("Taux de faux positifs")
plt.ylabel("Taux de vrais positifs")
plt.title("Courbe ROC en validation croisée")
plt.legend(loc="lower right")
plt.show()

# Entraînement final
rf.fit(X_scaled, y)

# Importance des variables
importances = rf.feature_importances_
plt.figure(figsize=(8,5))
sns.barplot(x=importances, y=features)
plt.title("Importance des variables (modèle final)")
plt.xlabel("Importance")
plt.ylabel("Variables")
plt.show()
# Données batch à prédire
data_brut = """
1 1 33.12 100 2.99 1 2 0 1 1 0
0 1 1.73 1.32 9.13 1 1 0 1 1 1
0 0 11.88 5.51 9.58 0 0 0 1 1 0
0 0 1.9 0.55 1.32 2 2 0 1 1 1
0 0 8.84 8.59 4.94 0 2 0 1 1 0
0 0 3.9 3.26 4.23 0 0 0 1 1 0
1 1 14.02 1.48 9.57 0 1 0 1 1 0
1 1 11.2 4.3 5.26 0 0 0 1 1 0
0 0 16.18 6.46 5.75 0 0 0 1 1 0
0 1 9.05 5.2 1.62 0 0 0 1 1 1
0 1 35.83 12.04 9.56 2 2 0 1 1 1
0 1 20.6 15 3.26 2 2 0 1 1 0
1 1 3.3 3.11 15.17 0 0 1 1 1 0
1 0 1.83 7.85 15 0 0 1 1 1 0
1 0 21.53 11.89 2.22 2 1 0 1 1 0
1 0 21.5 15 1.37 2 2 0 1 1 0
0 1 25.91 25.18 1.36 1 1 0 1 1 1
0 1 9.2 3.95 2.79 0 0 0 1 1 0
0 0 4.09 3.97 3.99 0 0 1 1 1 0
0 1 29.38 9.47 2.89 0 0 0 1 1 0
0 0 56.37 19.26 3.34 1 1 0 1 1 1
1 1 15.69 6.68 4.75 0 0 0 1 1 0
"""

lines = data_brut.strip().split("\n")
data = [list(map(float, line.split())) for line in lines]

# Extraction des 9 colonnes d'intérêt
data_9cols = [[row[i] for i in [0,1,2,3,4,5,6,7,10]] for row in data]

data_scaled = scaler.transform(data_9cols)

predictions = rf.predict(data_scaled)
probas = rf.predict_proba(data_scaled)
```

```
print("\n==== PRÉDICTIONS POUR LES NOUVEAUX PATIENTS ====")
for i, (pred, proba) in enumerate(zip(predictions, probas)):
    conf = proba[pred] * 100
    print(f"Patient {i+1} : Prédiction = {pred} "
          f"(Confiance = {conf:.2f}%) "
          f"(0 = azoospermie sévère, 1 = non azoospermie)")

# === AJOUT : comparaison avec vrais résultats ===

# Vrais résultats fournis (0 ou 1), dans l'ordre des patients testés
y_true = [1,0,1,1,0,0,0,0,0,0,0,1,1,0,0,0,1,0,1,0,0,0,1]
print(f"Nombre de prédictions : {len(predictions)}")
print(f"Nombre de vrais résultats : {len(y_true)}")
# Ajustement si nécessaire
min_len = min(len(predictions), len(y_true))
predictions = predictions[:min_len]
probas = probas[:min_len]
y_true = y_true[:min_len]

print("\n==== COMPARAISON PREDICTIONS / VRAIS RESULTATS ====")
for i, (pred, proba, vrai) in enumerate(zip(predictions, probas, y_true)):
    conf = proba[pred] * 100
    correct = "✓" if pred == vrai else "✗"
    print(f"Patient {i+1} : Prédiction = {pred} "
          f"(Confiance = {conf:.2f}%) "
          f"Vrai = {vrai} "
          f"{correct}")



cm = confusion_matrix(y_true, predictions)
plt.figure(figsize=(6,5))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False,
            xticklabels=['Prédit 0', 'Prédit 1'],
            yticklabels=['Réel 0', 'Réel 1'])
plt.xlabel("Classe prédite")
plt.ylabel("Classe réelle")
plt.title("Matrice de confusion : comparaison prédictions vs vrais résultats")
plt.show()
```