REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة

التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة أبي بكر بلقايد – تلمسان –

Université Aboubakr Belkaïd – Tlemcen –

Faculté de TECHNOLOGIE

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Génie Biomédical

Spécialité : Electronique et Maintenance Biomédicale

Par : HAMIANI Hichem Mounssif et HADJIAT Mohamed Ghouti

Sujet

UN SYSTÈME D'AIDE AU DIAGNOSTIC DU CANCER DU SEIN (CLASSIQUE)

Soutenu publiquement, le 29 / 06 /2025, devant le jury composé de :

Mme HABIBES Naima	MAA	Université de Tlemcen	Présidente
M.HADJ Ahmed Ismail	MCB	Université de Tlemcen	Examinateur
M. BENALI Radhwane	MCA	Université de Tlemcen	Encadreur

Année universitaire : 2024 / 2025

Remerciements

Nous tenons tout d’abord à exprimer notre profonde gratitude à Allah, pour nous avoir accordé la force, la patience et la persévérance nécessaires à l’accomplissement de ce travail.

Nous adressons nos sincères remerciements à Monsieur BENALI RADHWANE, notre encadrant, pour son accompagnement tout au long de ce projet. Sa disponibilité, ses conseils avisés, ainsi que sa rigueur scientifique ont été d’une grande aide et ont grandement contribué à l’aboutissement de ce mémoire. Nous lui exprimons toute notre reconnaissance pour la confiance qu’il nous a accordée.

Nous remercions également toutes les personnes qui nous ont soutenus de près ou de loin durant ce parcours. Leurs encouragements, leur bienveillance et leur appui constant ont été précieux tout au long de cette aventure académique.

À toutes celles et ceux qui ont, d’une manière ou d’une autre, contribué à la réalisation de ce mémoire, nous exprimons notre gratitude la plus sincère.

Mot de l'étudiant

Issu d'une formation en électronique et aujourd'hui étudiant en Master 2 Électronique et Maintenance Biomédicale, j'ai entamé mon parcours avec la conviction que la technologie, lorsqu'elle est bien dirigée, peut devenir une alliée précieuse au service de la vie humaine. Mon intérêt initial pour les systèmes techniques s'est peu à peu transformé en une quête de sens, m'orientant naturellement vers le domaine biomédical, là où l'innovation technologique rencontre les besoins urgents de la santé.

Lors de mon dernier stage au sein de l'entreprise EURL GTM, spécialisée dans les solutions d'imagerie médicale, j'ai occupé le poste d'ingénieur biomédical – application échographie. J'ai eu l'opportunité d'accompagner les professionnels de santé dans l'installation et l'utilisation d'équipements d'échographie. Ce fut une immersion directe dans la réalité clinique, où chaque image, chaque réglage, peut avoir un impact décisif sur un diagnostic, sur une vie.

C'est au contact de ces scènes humaines et techniques que j'ai pris conscience de la puissance de l'intelligence artificielle appliquée à l'imagerie médicale. J'ai vu des examens, autrefois longs et fastidieux, être raccourcis grâce à l'IA. J'ai compris à quel point cette technologie, loin de remplacer l'expertise humaine, pouvait la sublimer : en aidant à la classification des tissus, à l'estimation des volumes, ou à l'identification précoce d'anomalies. L'avenir de la médecine, je l'ai perçu là, dans cette synergie entre l'intelligence humaine et l'intelligence artificielle.

C'est dans ce contexte que le thème de mon projet de fin d'études s'est imposé avec évidence : concevoir un système d'aide à la détection du cancer du sein. Plus qu'un projet académique, il s'agit pour moi d'une réponse technique à une urgence humaine. C'est aussi une manière d'exprimer ma vision de l'ingénieur biomédical : un professionnel capable de transformer la complexité technologique en solutions concrètes, accessibles et utiles pour les praticiens comme pour les patients.

Ce mémoire est le fruit de cette ambition, nourrie par l'expérience, le terrain, et une volonté sincère de contribuer, à mon échelle, à l'amélioration des pratiques médicales.

HAMIANI Hichem Mounssif

Ce mémoire marque l'aboutissement d'un parcours universitaire riche en apprentissages et en expériences pratiques. Il m'a permis de mettre en application les connaissances acquises tout au long de ma formation et de développer une vision plus concrète de leur utilité dans un cadre professionnel.

Au cours des deux dernières années, ma spécialisation en électronique et maintenance biomédicale m'a permis de consolider mes bases en électronique tout en découvrant les spécificités des technologies utilisées dans le domaine médical. J'ai appris à comprendre le fonctionnement des dispositifs biomédicaux, à diagnostiquer les pannes, à assurer leur maintenance, et à respecter les normes de sécurité en milieu hospitalier. Cette formation allie rigueur scientifique, sens de l'observation et compétences techniques, et m'a préparée à intervenir efficacement dans des environnements sensibles.

J'ai également eu l'opportunité d'effectuer deux stages significatifs. Le premier au Centre d'imagerie Abelali, où j'ai découvert de près l'environnement médical et ses exigences. Le second, au sein de l'entreprise EURL MALI Plus à Alger, m'a permis d'acquérir une expérience sur le terrain dans la maintenance et l'installation de matériel d'anesthésie, notamment des ventilateurs artificiels et des moniteurs.

Ce travail m'a ainsi offert l'occasion de renforcer mes compétences, de gagner en autonomie, et de mieux me préparer aux exigences du monde professionnel.

HADJIAT Mohamed Ghouti

Résumé

Ce mémoire est consacré à la classification automatique du cancer du sein, un enjeu majeur pour l'amélioration du diagnostic médical. Structuré en trois chapitres, il présente d'abord les aspects médicaux du cancer du sein, puis détaille les fondements des algorithmes d'intelligence artificielle, notamment le SVM et la forêt aléatoire. Une étude expérimentale est ensuite menée sur une base de données clinique, avec une préparation rigoureuse des données et une comparaison des performances des modèles.

L'objectif principal est de proposer un outil d'aide à la décision médicale fiable et automatisé, capable de réduire les erreurs de diagnostic, notamment les faux négatifs, en s'appuyant sur des approches d'apprentissage supervisé. Ce travail met en lumière le potentiel de l'IA dans le domaine biomédical et ouvre la voie à des solutions innovantes et accessibles.

Mots clés : Classification automatique, Cancer du sein, Intelligence artificielle, Machine learning, SVM, Forêt aléatoire, Diagnostic assisté.

Abstract

This thesis focuses on the automatic classification of breast cancer, a major challenge in improving medical diagnosis. Structured in three chapters, it first presents the medical background of breast cancer, then explains the fundamentals of artificial intelligence algorithms, particularly Support Vector Machines (SVM) and Random Forest. An experimental study is conducted using a clinical dataset, with careful data preprocessing and a performance comparison of the models.

The main objective is to develop a reliable and automated decision support tool capable of reducing diagnostic errors, especially false negatives, using supervised learning techniques. This work highlights the potential of AI in the biomedical field and opens the door to innovative and accessible healthcare solutions.

Keywords: Automatic classification, Breast cancer, Artificial intelligence, Machine learning, SVM, Random Forest, Computer-aided diagnosis.

ملخص

تُركّز هذه المذكرة على موضوع التصنيف الآلي لسرطان الثدي، والذي يُعدّ من التحديات الكبرى في سبيل تحسين دقة التشخيص الطبي. تنقسم المذكرة إلى ثلاثة فصول، يتناول الفصل الأول الجوانب الطبية الأساسية المتعلقة بسرطان الثدي، بينما يشرح الفصل الثاني المبادئ النظرية لخوارزميات الذكاء الاصطناعي، مع التركيز على آلات الدعم الناقل وخوارزمية الغابة العشوائي. أما الفصل الثالث، فيتضمن دراسة تجريبية تعتمد على قاعدة بيانات سريرية، تتضمن معالجة منهجية للبيانات ومقارنة دقيقة بين أداء النماذج المدروسة.

يتمثل الهدف الأساسي من هذا العمل في تطوير أداة دعم قرار موثوقة وآلية، تساعد على تقليل أخطاء التشخيص، لا سيما حالات النتائج السلبية الكاذبة، وذلك باستخدام تقنيات التعلم الخاضع للإشراف. ويبرز هذا المشروع الإمكانيات الكبيرة التي يوفرها الذكاء الاصطناعي في المجال الطبي الحيوي، ويمهّد الطريق نحو حلول صحية مبتكرة وأكثر توفراً.

Table des Matières

Table des illustrations

Liste des tableaux

Introduction Générale

Le cancer du sein représente aujourd'hui l'une des principales préoccupations de santé publique à l'échelle mondiale. En Algérie, il constitue la première cause de mortalité par cancer chez la femme, touchant chaque année un grand nombre de patientes, souvent à un stade avancé. Ce constat alarmant s'explique par plusieurs facteurs : un diagnostic tardif, l'absence de dépistage systématique et les difficultés d'accès aux moyens d'imagerie médicale avancés. Ces obstacles freinent considérablement une prise en charge rapide et efficace.

Face à cette situation, l'intelligence artificielle (IA) se présente comme une technologie d'avenir, notamment grâce à ses capacités d'analyse automatisée et rapide de grandes quantités de données médicales. Plus particulièrement, les méthodes de classification automatique permettent de développer des systèmes capables d'assister les professionnels de santé dans la détection et l'évaluation des tumeurs mammaires. Ces outils peuvent améliorer la précision des diagnostics, réduire le temps d'analyse, et offrir des résultats plus homogènes, tout en limitant les erreurs humaines.

1. Problématique

Dans de nombreux contextes cliniques, en particulier ceux présentant des ressources limitées, la fiabilité et la rapidité du diagnostic du cancer du sein restent des défis majeurs. Le recours à des solutions automatisées pourrait réduire la charge de travail du personnel médical, mais ces solutions doivent être techniquement solides, cliniquement pertinentes, et capables de minimiser les erreurs critiques, notamment les faux négatifs.

Dès lors, la problématique centrale de ce travail peut être formulée ainsi :
Comment concevoir un système de classification automatique du cancer du sein, à partir de données réelles, capable de distinguer de manière fiable les tumeurs bénignes des tumeurs malignes, tout en s'intégrant dans un cadre clinique réaliste ?

2. Méthodologie

La méthodologie adoptée dans ce projet repose sur l'utilisation d'algorithmes d'apprentissage automatique (machine learning), appliqués à la base de données Wisconsin Breast Cancer, largement utilisée dans les recherches biomédicales.
Elle se décline en plusieurs étapes principales :

- Prétraitement des données : nettoyage, conversion et sélection des variables pertinentes
- Application des modèles : implémentation des algorithmes SVM (Support Vector Machine) et Random Forest
- Évaluation des performances : à l'aide de métriques comme la précision, le rappel, la F-mesure et la matrice de confusion

- Comparaison et analyse : identification du modèle le plus performant et des combinaisons de variables optimales

3. Objectifs

Ce projet vise deux objectifs principaux :

- Technique : développer un système de classification automatisée capable de traiter des données médicales réelles, avec une performance fiable dans la détection des tumeurs mammaires
- Clinique : proposer un outil d'aide à la décision pouvant renforcer le dépistage précoce, en particulier dans les environnements où le personnel qualifié ou les équipements spécialisés sont insuffisants
- Un objectif secondaire mais essentiel est de réduire les erreurs de diagnostic, notamment les faux négatifs, qui représentent un risque important pour la santé des patientes.

4. Plan du mémoire

Le contenu de ce mémoire est organisé en trois chapitres :

- Chapitre 1 : Présente les généralités sur le cancer du sein, incluant l'anatomie, les types de tumeurs, les symptômes, les méthodes de diagnostic et le contexte en Algérie
- Chapitre 2 : Décrit les fondements théoriques des algorithmes de classification automatique, avec un focus sur le machine learning, les SVM et les forêts aléatoires
- Chapitre 3 : Détaille l'étude expérimentale réalisée sur des données médicales, en appliquant les modèles et en évaluant leurs performances pour la classification des tumeurs

Chapitre 1
Généralités sur le cancer du sein

1. Introduction

Ce chapitre est consacré aux fondements médicaux du cancer du sein, dans le but de fournir une compréhension Simplifié de cette pathologie qui constitue aujourd'hui un enjeu majeur de santé publique à l'échelle mondiale. Le cancer du sein représente le type de cancer le plus fréquemment diagnostiqué chez la femme et demeure l'une des principales causes de mortalité féminine liée au cancer.

L'analyse des mécanismes biologiques, des facteurs de risque, ainsi que des méthodes actuelles de détection et de diagnostic, est essentielle pour appréhender les défis posés par cette maladie.

2. Le cancer du sein en Algérie

Le cancer du sein est le cancer le plus fréquent chez la femme algérienne. Il représente également une cause majeure de mortalité par cancer dans cette population [1]. Les données récentes du Registre des Tumeurs d'Alger indiquent une augmentation continue du nombre de cas. En 2021, l'incidence brute était de 93,1 cas pour 100 000 femmes, contre 78,8 en 2020 [1]. Ce cancer représente 41,5 % de l'ensemble des cancers féminins enregistrés dans la wilaya d'Alger.

L'âge moyen au diagnostic est de 51,3 ans, avec une concentration des cas dans la tranche d'âge 45–54 ans. Cette moyenne est inférieure à celle observée dans les pays européens, où elle dépasse souvent 60 ans [2]. Cette différence suggère une atteinte plus précoce dans la population algérienne.

Sur le plan histologique, le carcinome canalaire infiltrant est la forme la plus fréquente, représentant 79,2 % des cas [1]. La majorité des diagnostics sont posés à un stade avancé, ce qui impacte négativement le pronostic et complique la prise en charge. Selon l'Organisation mondiale de la Santé, un dépistage précoce permettrait de guérir plus de 90 % des cancers du sein détecté à un stade initial [2]. En l'absence d'un programme national de dépistage organisé, de nombreuses patientes consultent tardivement, réduisant ainsi les chances de traitement curatif.

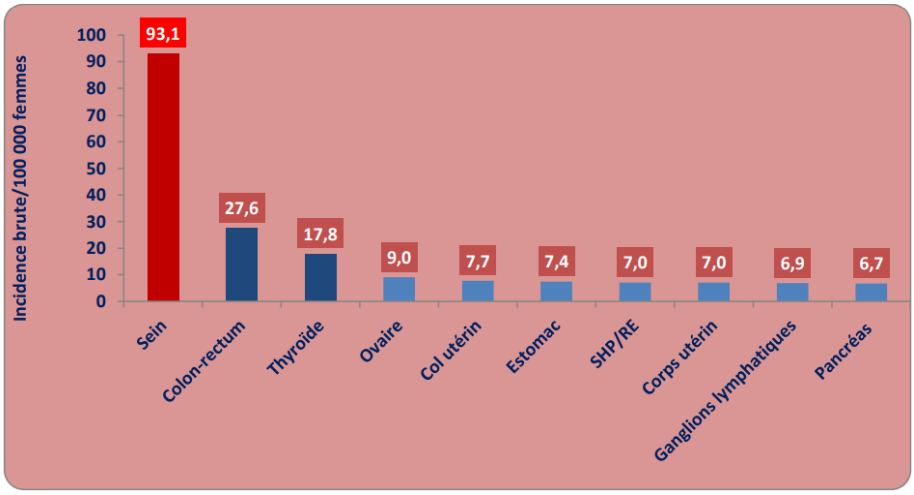

Figure 1 - Localisations cancéreuses les plus fréquentes chez les femmes

3. Anatomie du sein

1. Localisation et forme

Le sein est une structure arrondie située à l'avant du thorax, entre la clavicule et la base du sternum. Il peut légèrement s'étendre vers la région de l'aisselle. Bien qu'il repose sur le muscle grand pectoral, il n'en fait pas partie ; il est séparé de celui-ci par une fine couche de tissu [3].

2. Composants principaux

Tissu adipeux

C'est le tissu gras qui compose la majeure partie du sein. Il détermine en grande partie son volume et sa forme, qui peuvent varier selon l'âge, le poids et les hormones.

Tissu glandulaire

Il est formé de 15 à 20 régions appelées lobes. Chacun de ces lobes est constitué de petits groupes appelés lobules, qui sont capables de produire du lait.

Canaux galactophores

Ces fins conduites permettent de transporter le lait depuis les lobules jusqu'au mamelon, où il peut être libéré lors de l'allaitement [3].

3. Peau et mamelon

Aréole

Partie colorée entourant le mamelon. Elle contient de petites glandes qui hydratent et protègent la peau, en particulier pendant l'allaitement.

Mamelon

Zone saillante située au centre de l'aréole, par laquelle passent les canaux lactifères. Il est particulièrement sensible, riche en terminaisons nerveuses [3].

4. Structures de soutien

Ligaments de soutien

Ce sont des bandes fibreuses internes qui relient le sein à la peau et aux muscles profonds. Ils participent à maintenir sa forme et sa fermeté.

Tissu conjonctif

Ce tissu de soutien enveloppe les différentes parties internes du sein (lobes, canaux) et contribue à sa cohésion et sa stabilité [3].

5. Vascularisation et drainage

Circulation sanguine

Le sein est alimenté en sang par des vaisseaux artériels et veineux qui assurent l'apport en nutriments et en oxygène nécessaires à son bon fonctionnement.

Système lymphatique

Un réseau de vaisseaux permet d'évacuer les déchets et de participer à la défense immunitaire. Les ganglions situés sous l'aisselle jouent un rôle central dans ce processus, en particulier en cas d'infection ou de pathologie [3].

6. Fonction principale

- Chez la femme, le sein produit du lait pour nourrir le nouveau-né.
- Les lobules fabriquent le lait (sous l'action des hormones).
- Les canaux le transportent jusqu'au mamelon, où il est évacué lors de l'allaitement [3].

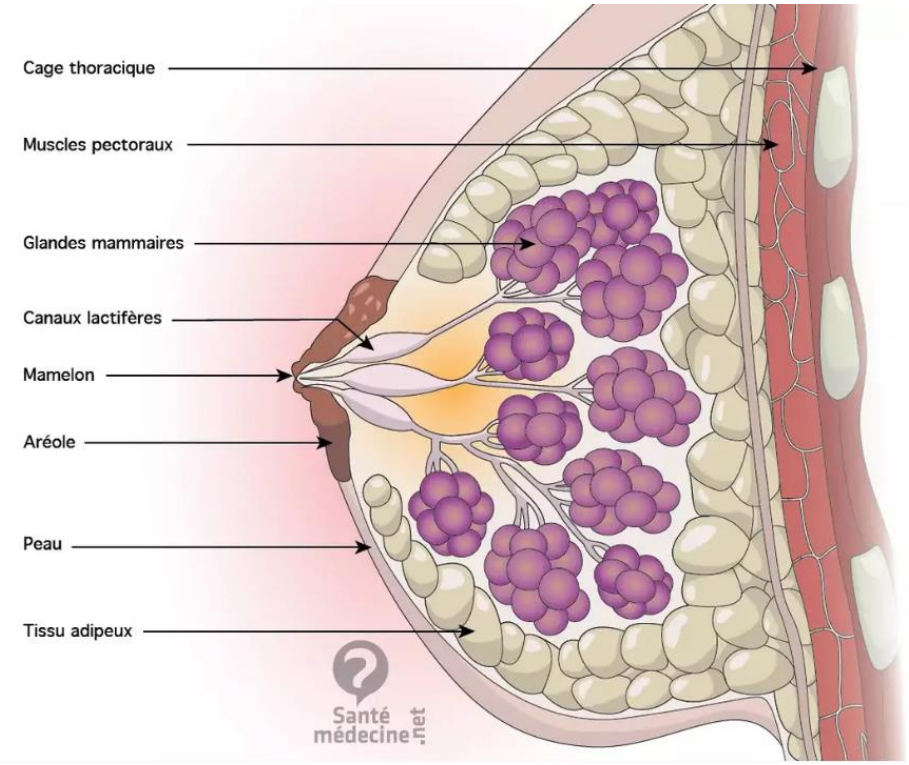

Figure 2 - Schéma en coupe anatomique du sein

4. Le cancer du sein

1. Définition et fonctionnement

Définition du cancer

Le cancer est une maladie causée par un dérèglement du cycle cellulaire. Le corps humain est constitué de plus de deux cents types de cellules, chacune ayant une fonction spécifique, comme les cellules musculaires, nerveuses ou osseuses [4].

Cellule cancéreuse

Une cellule devient cancéreuse lorsqu'elle subit des modifications anormales qu'elle n'arrive plus à corriger. Elle perd son contrôle de croissance, se multiplie de façon incontrôlée et finit par former une tumeur maligne [4].

Tumeur bénigne vs maligne

Une tumeur bénigne est une masse de cellules anormales, mais non invasive. Elle reste localisée et ne se propage pas à d'autres parties du corps. En revanche, une tumeur maligne peut envahir les tissus voisins et se développer à distance [4].

Cancer du sein

Le cancer du sein est une forme de tumeur maligne qui se développe dans les tissus mammaires. Il concerne environ une femme sur onze au cours de sa vie [4].

Formes de cancer du sein

Le cancer peut être dit "in situ", ce qui signifie qu'il reste localisé à l'endroit où il est apparu, sans se propager aux tissus autour.

Dans d'autres cas, le cancer devient "infiltrant" : cela veut dire que les cellules cancéreuses commencent à envahir les tissus voisins. Elles peuvent atteindre les ganglions axillaires, qui sont de petits organes situés sous les aisselles et qui jouent un rôle dans le système de défense du corps.

Si la maladie progresse, certaines cellules cancéreuses peuvent quitter la tumeur et circuler dans le corps en passant par le sang ou par la lymphe, un liquide qui transporte les cellules de défense. Elles peuvent alors aller s'installer dans d'autres parties du corps, comme les poumons, le foie ou les os, et y former de nouvelles tumeurs. Ce phénomène s'appelle une métastase et correspond à un stade plus avancé du cancer [4].

2. Facteurs de risque

Les facteurs de risque se divisent en non modifiables et modifiables :

Facteurs non modifiables

- Âge et sexe : être une femme et vieillir (risque plus élevé après la ménopause).
- Prédisposition génétique/familiale : antécédents personnels ou familiaux de cancer du sein.
- Facteurs hormonaux/reproductifs : premières règles précoces, ménopause tardive ou âge tardif au premier enfant [5].

Facteurs modifiables

- Hormones sexuelles exogènes : utilisation prolongée d'hormonothérapie.
- Mode de vie : surpoids/obésité, sédentarité, consommation d'alcool et alimentation déséquilibrée.
- Expositions : radiations ionisantes thoraciques (ex. traitement contre certaines pathologies) et environnement sont suspectés comme facteurs contributifs [5].

3. Types histologiques

Il existe plusieurs types de cancer du sein, mais les deux plus fréquents sont :

Le carcinome canalaire infiltrant

- C'est le plus courant. Il commence dans les canaux qui transportent le lait à l'intérieur du sein.

Le carcinome lobulaire infiltrant

- Un peu moins fréquent, il débute dans les zones du sein qui produisent le lait.

Ces deux types représentent la grande majorité des cancers du sein [6].

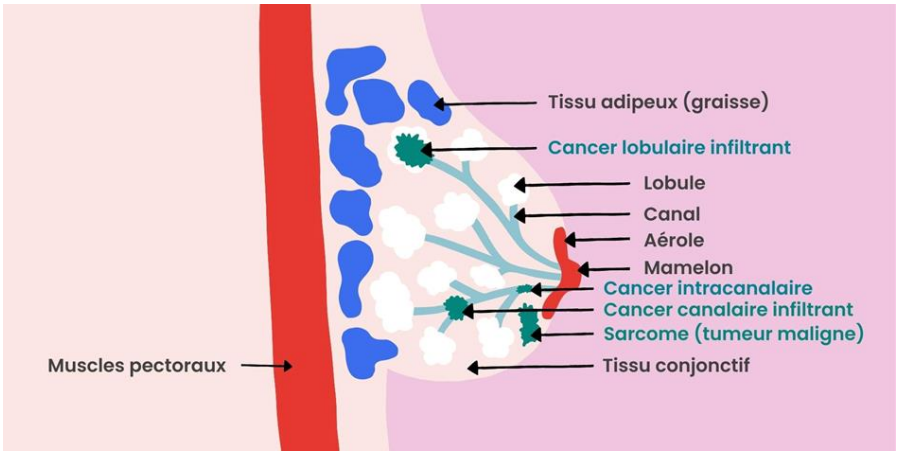

Figure 3 - Schéma illustratif des différents types histologiques du cancer du sein [7]

4. Symptomatologie

Le cancer du sein peut être silencieux aux premiers stades (découvert lors d'un dépistage radiologique). Lorsque des symptômes surviennent, les plus fréquents sont [7] :

Nodule mammaire

- Masse ferme ou dure, souvent fixée aux plans profonds, non régressant pendant le cycle menstruel.

Modification du mamelon

- Inversion soudaine du mamelon ou rétraction, écoulement séreux ou hémorragique spontané.

Changements de la peau

- Œdème cutané avec aspect « peau d'orange », rougeur, ulcération ou érosion de la peau mammaire (forme inflammatoire).

Adénopathie axillaire

- Masse ou durcissement dans le creux de l'aisselle (ganglions palpables), parfois révélateur.

Autres signes possibles

- Modification de la taille/forme du sein, sensation de « chaud » localisé. Dans les stades avancés (métastases), on peut voir des signes systémiques (fatigue, amaigrissement).

Formes asymptomatiques

- De nombreux cancers lobulaires ou carcinomes in situ restent longtemps sans signe clinique et ne sont détectés que par la mammographie de dépistage.

5. Diagnostic

Le diagnostic repose sur l'imagerie et l'analyse histologique [8] :

Mammographie

C'est un examen radiologique des seins, utilisant des rayons X à faible dose pour visualiser l'intérieur du tissu mammaire. Grâce à la compression du sein entre deux plaques, on obtient des images plus nettes et à moindre irradiation.

Mammographie de dépistage

Il s'agit de la même mammographie que précédemment, mais réalisée de façon **préventive** (par exemple, tous les deux ans chez les 50–74 ans), afin de repérer **tôt** des lésions encore invisibles à l'autopalpation.

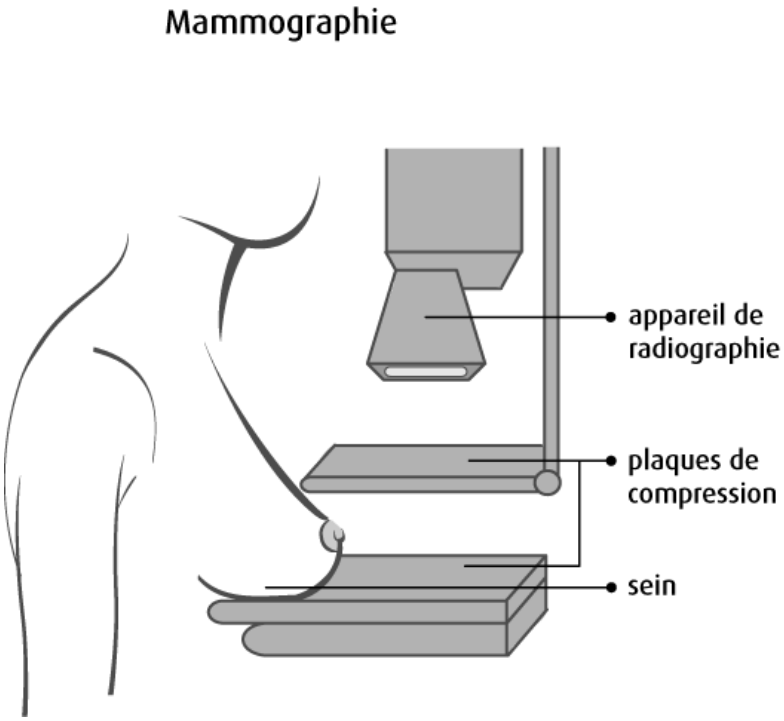

Figure 4 - visualisation typique d'un sein comprimé lors du cliché

Échographie mammaire

Technique d'imagerie non irradiante, l'échographie mobilise des ultrasons pour générer des images en temps réel. Elle sert souvent d'examen complémentaire à la mammographie, notamment pour distinguer une masse solide.

Biopsie

Il s'agit d'un prélèvement de tissu mammaire réalisé à l'aide d'une aiguille insérée à travers la peau, sous anesthésie locale et guidé par imagerie (radiographie ou échographie), afin d'analyser en laboratoire si la lésion est bénigne ou maligne.

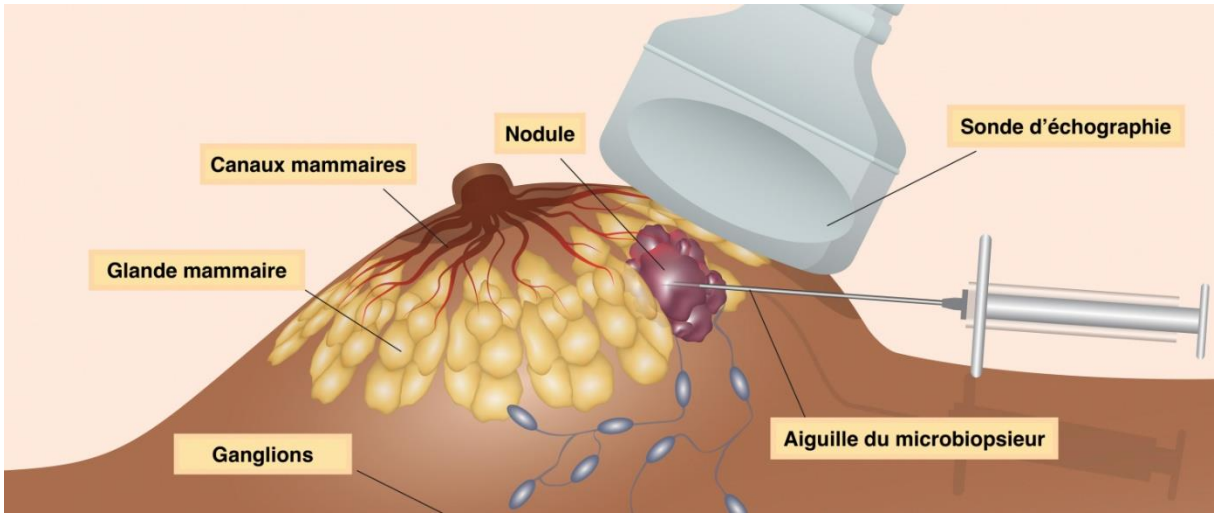

Figure 5 - Biopsie mammaire guidée par échographie

6. Analyse anatomopathologique d'une biopsie mammaire

L'analyse d'un prélèvement de tissu mammaire (biopsie) passe par plusieurs étapes précises. Ces étapes permettent de préserver, préparer et observer le tissu au microscope afin de détecter d'éventuels signes de cancer. C'est une méthode essentielle en médecine.

Fixation : préserver les cellules

Dès le prélèvement, le tissu est placé dans une solution appelée formol tamponné à 10 %. Cette étape permet de stabiliser les cellules et d'éviter leur dégradation. Le formol fige les protéines pour garder la forme des cellules aussi proche que possible de leur état naturel [10]. En général, cette fixation dure entre 6 et 24 heures selon la taille du fragment.

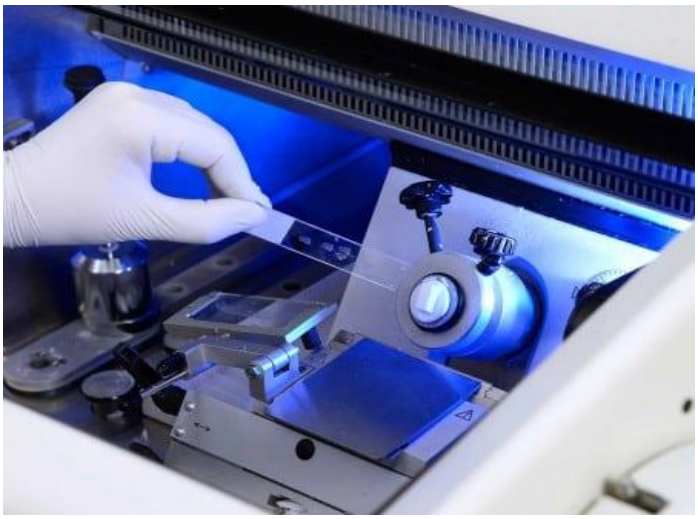

Figure 6 - Fixation du prélèvement

Inclusion en paraffine : rendre le tissu manipulable

Après la fixation, le tissu est déshydraté (l'eau est retirée avec des bains d'alcool), puis plongé dans du xylène, avant d'être infiltré par de la paraffine chaude (vers 60 °C). Une fois durci, ce bloc de paraffine permet de manipuler facilement le tissu [11].

Coupe histologique : obtenir des tranches très fines

Le bloc est placé dans un microtome, un appareil qui découpe des tranches très fines (4 à 5 micromètres d'épaisseur). Ces coupes sont ensuite déposées sur des lames de verre pour être observées au microscope [12].

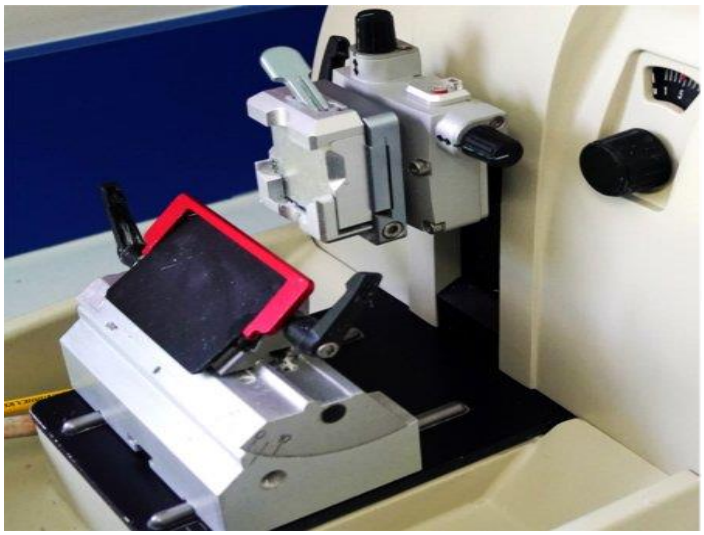

Figure 7 - Bloc de paraffine

Coloration H&E : faire ressortir les détails des cellules

Pour distinguer les différentes structures au microscope, on colore les lames avec la technique H&E (Hématoxyline–Éosine) :

- L'hématoxyline colore les noyaux en bleu-violet (où se trouve l'ADN),
- L'éosine colore les tissus environnants en rose [13].

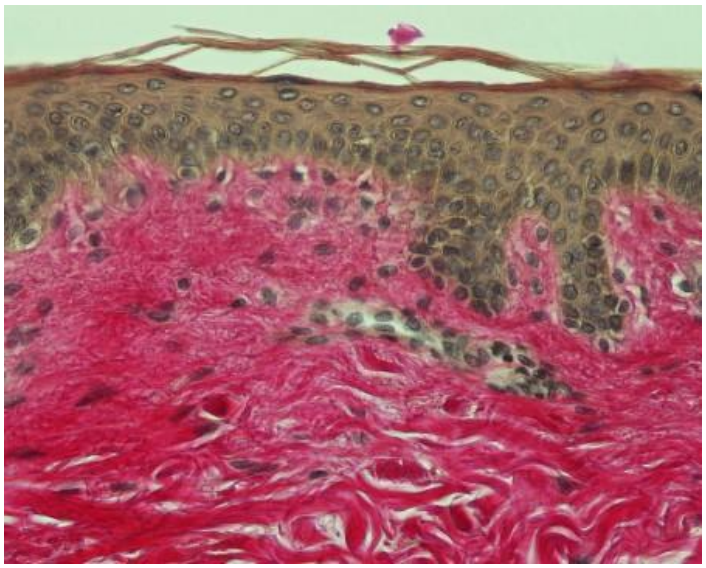

Figure 8 - Coupe histologique d'un tissu mammaire coloré H&E

Observation au microscope : identifier les signes de cancer

- Le pathologiste examine la lame contenant l'échantillon de tissu, préalablement colorée, à l'aide d'un microscope optique (un appareil qui permet de voir les cellules à très fort grossissement).

- Il observe la morphologie des cellules, c'est-à-dire leur forme, leur taille, ainsi que l'aspect des noyaux (la partie centrale de la cellule qui contient le matériel génétique), en étudiant leur densité et leur régularité.
- La disposition des cellules dans le tissu est également analysée. Cela permet d'évaluer l'organisation générale des cellules entre elles, qu'on appelle l'architecture tissulaire.
- Certains éléments peuvent indiquer la présence d'un cancer, comme des divisions cellulaires anormales (appelées mitoses atypiques), des noyaux de forme inhabituelle (appelés atypies nucléaires), ou encore une désorganisation de la structure normale du tissu [14].
- Cette analyse précise permet de faire la différence entre une tumeur bénigne (non cancéreuse) et une lésion maligne (cancéreuse). Le diagnostic ainsi obtenu, appelé diagnostic histologique, est essentiel pour choisir le traitement adapté et décider de la suite de la prise en charge du patient.

5. Conclusion

L'étude menée dans ce chapitre, centrée sur les aspects médicaux et anatomiques, nous a permis d'approfondir notre compréhension du cancer du sein. Nous avons également analysé les méthodes classiques de diagnostic, encore couramment utilisées par les médecins de manière manuelle. Ces éléments constituent une base essentielle pour aborder, dans les chapitres suivants, les innovations technologiques susceptibles d'améliorer la détection et la prise en charge de cette pathologie.

Dans ce contexte, il devient crucial de développer des outils d'aide à la décision capables d'analyser automatiquement les données médicales issues des examens de dépistage. L'intégration de solutions informatiques intelligentes permettrait non seulement de soutenir les professionnels de santé dans leur prise de décision, mais aussi de garantir une meilleure objectivité et reproductibilité des résultats.

L'objectif de ce projet est donc d'explorer et d'évaluer différentes approches algorithmiques issues de l'intelligence artificielle afin de proposer un système de classification automatique fiable, basé sur des données cliniques réelles. Ce système vise à distinguer les cas bénins des cas malins, dans le but de faciliter un dépistage plus rapide, plus précis et mieux adapté aux contraintes du domaine médical.

Chapitre 2 Théorie des algorithmes de classification

1. Introduction

La classification des tumeurs mammaires constitue une étape fondamentale dans le diagnostic du cancer du sein. En effet, une identification précise du type de tumeur permet d'adapter les choix thérapeutiques aux caractéristiques biologiques de la lésion. Dans cette optique, il est essentiel de disposer d'une méthode de classification à la fois rigoureuse, fiable et applicable à différentes situations cliniques.

Ce chapitre présente le cadre théorique utilisé pour la classification des tumeurs mammaires à partir de données médicales. L'objectif est d'établir une approche claire, reproductible et suffisamment flexible pour être appliquée dans divers contextes expérimentaux ou cliniques.

2. Contexte et objectifs de la classification automatique

Dans ce contexte, la classification automatique des tumeurs mammaires s'impose comme une solution innovante et complémentaire aux méthodes traditionnelles. Elle consiste à utiliser des algorithmes informatiques, basés sur l'intelligence artificielle et l'apprentissage automatique (machine learning), capables d'apprendre à partir de données existantes et de classer de nouvelles observations selon des critères préalablement définis.

Les objectifs spécifiques de la classification automatique dans ce contexte sont les suivants :

- Améliorer la précision du diagnostic grâce à des modèles capables d'identifier des schémas complexes dans les données.
- Réduire le temps de traitement en automatisant l'analyse et l'interprétation des données.
- Standardiser les résultats en limitant l'influence des variations humaines.
- Fournir un outil d'aide à la décision pour soutenir les cliniciens dans la détection et le suivi des tumeurs mammaires.
- Faciliter le traitement de grands volumes de données, en particulier dans le cadre de bases de données cliniques ou d'images médicales.

1. L'intelligence artificielle

L'intelligence artificielle (IA) désigne la capacité d'un système informatique à réaliser des tâches qui requièrent habituellement une intervention humaine, comme l'analyse, la décision ou l'adaptation à une situation. Il s'agit de programmes capables d'apprendre ou de s'ajuster à partir de données, comme les assistants numériques ou les systèmes de diagnostic médical.

L'IA regroupe un ensemble de disciplines visant à créer des outils capables de simuler certaines formes d'intelligence humaine, dans des domaines variés [15].

2. Apprentissage automatique (Machine Learning)

L'apprentissage automatique est une branche de l'IA qui se concentre sur des techniques permettant aux ordinateurs d'apprendre à partir de données. Le système améliore ses performances sur une tâche donnée sans qu'on lui donne de règles précises à chaque fois.

Il utilise des outils statistiques pour détecter des motifs dans les données (images, chiffres, signaux…) et s'en servir pour faire des prédictions sur de nouveaux cas [16].

3. Apprentissage supervisé et non supervisé

En apprentissage supervisé, le système apprend à partir d'exemples annotés, c'est-à-dire pour lesquels la réponse correcte est connue. Cela permet à l'algorithme d'apprendre la correspondance entre les données et leur catégorie (par exemple : tumeur bénigne ou maligne) [17].

À l'inverse, l'apprentissage non supervisé consiste à découvrir des structures cachées dans des données non annotées. L'algorithme tente d'identifier des regroupements naturels, par similarité, sans qu'on lui donne de catégories à l'avance [18]. Cette méthode est utile pour explorer des données inconnues ou détecter des sous-groupes dans une population [18].

4. Apprentissage profond (Deep Learning)

Le deep learning est une méthode avancée de l'apprentissage automatique, qui s'appuie sur des réseaux de neurones artificiels constitués de nombreuses couches. Chaque couche extrait des informations de plus en plus complexes à partir des données d'entrée.

Par exemple, dans le traitement d'images, les premières couches détectent des formes simples (lignes, angles), tandis que les couches suivantes reconnaissent des objets plus complexes. Ce type d'architecture est particulièrement efficace pour des tâches comme la reconnaissance visuelle ou vocale [19].

Le succès du deep learning repose sur deux éléments : la disponibilité de grandes quantités de données et la puissance de calcul moderne. Toutefois, pour des données structurées de taille modeste, des méthodes classiques comme les forêts aléatoires ou les modèles linéaires restent plus rapides, plus simples à entraîner et souvent aussi efficaces [19].

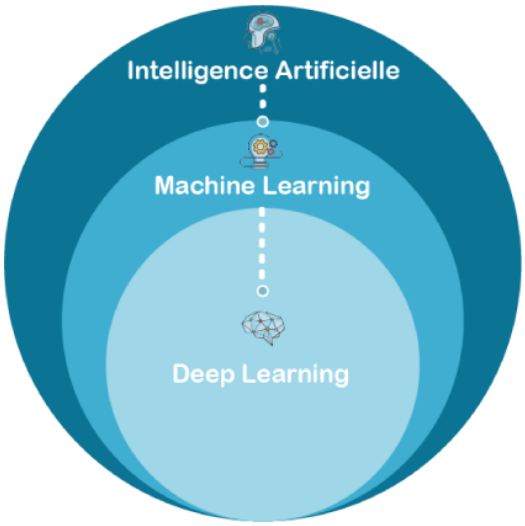

Figure 9 - Machine Learning vs Deep Learning

5. Classification

La classification est une tâche d'apprentissage consistant à attribuer une catégorie à chaque élément analysé. Dans le cas des tumeurs mammaires, elle permet par exemple de déterminer si une image correspond à une tumeur bénigne ou maligne.

La classification supervisée s'appuie sur des exemples annotés : le modèle apprend à reconnaître les caractéristiques associées à chaque classe [17].

La classification non supervisée, ou clustering, cherche à regrouper les données selon leurs similarités sans indication préalable. Elle est utile pour révéler des structures ou des groupes cachés dans des ensembles complexes [20].

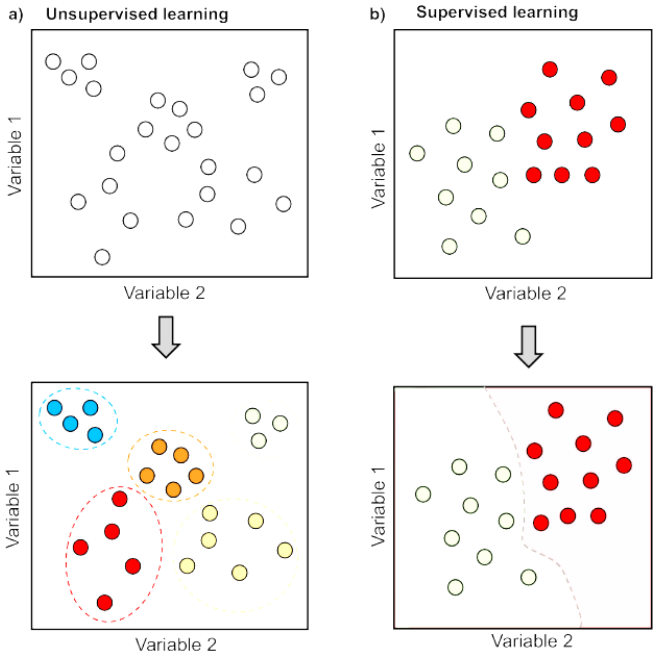

Figure 10 - Apprentissage supervisé vs non supervisé

6. Définition d'un modèle de machine learning

Un modèle de machine learning (ou d'apprentissage automatique) est une fonction mathématique construite à partir d'exemples, capable d'identifier des motifs dans les données et de faire des prédictions sur des cas nouveaux [21].

Entrée/Sortie

Il associe des données d'entrée (comme des images, des séries de nombres ou des signaux) à une sortie cible (catégories ou valeurs numériques), en ajustant ses paramètres internes pendant une phase d'apprentissage supervisé [22].

Apprentissage automatique

Contrairement aux programmes classiques qui suivent des instructions fixes, ce type de modèle découvre automatiquement les relations présentes dans les données, sans que chaque règle de traitement soit explicitement codée [23].

Capacité de généralisation

Une fois formé, le modèle peut traiter des données jamais vues auparavant. Cela le rend apte à résoudre des tâches variées telles que la classification d'images, la reconnaissance vocale ou la détection d'anomalies [24].

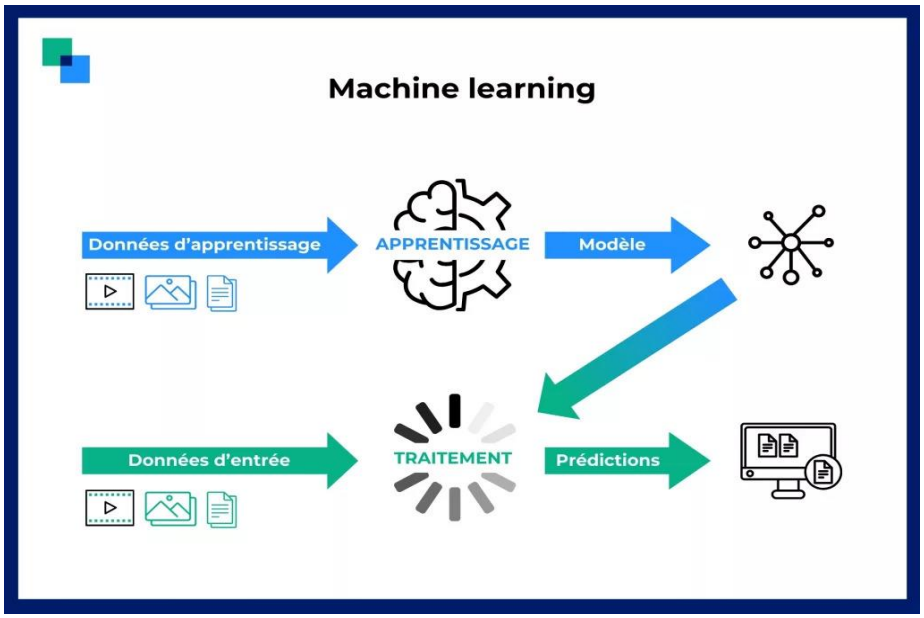

Figure 11 - fonctionnement du modèle de machine learning

7. Principaux types de modèles de machine learning

Les modèles supervisés se distinguent par leur structure et leur manière d'interpréter les données. Voici les grandes familles les plus utilisées :

Modèles linéaires simples

Ces modèles supposent que la sortie est obtenue à partir d'une combinaison linéaire des variables d'entrée. Par exemple, en régression linéaire, on ajuste une droite (ou un plan) qui s'approche au mieux des données. Cela permet une interprétation directe des relations entre variables, avec un entraînement rapide. En classification, ils sont à la base de la régression logistique ou de fonctions de score simples [25].

Arbres de décision

Les arbres de décision divisent progressivement les données en sous-groupes à l'aide de conditions hiérarchiques. Chaque nœud teste une variable, et chaque branche conduit à une prédiction en fonction des réponses [26]. Cette approche, facile à comprendre, peut être limitée en précision lorsqu'elle est utilisée seule, d'où le recours fréquent à des ensembles d'arbres, comme les forêts aléatoires.

Réseaux de neurones artificiels

Inspirés du cerveau humain, les réseaux de neurones sont constitués de plusieurs couches connectées. Ils permettent d'apprendre des relations complexes entre variables. Avec une ou deux couches intermédiaires, on parle de réseaux peu profonds, adaptés à des tâches de classification ou de régression [27].

3. Machines à vecteurs de supports

Les SVM (en anglais : « Support Vector Machines ») est une méthode qui a été créée par deux scientifiques : Vladimir Vapnik et Alexey Chervonenkis pendant les années 60. Par la suite, les SVM sont devenues très populaire pour résoudre les problèmes de classification [28].

Principe de fonctionnement

Ces modèles d'apprentissage supervisé sont utilisés pour classifier des données en identifiant un hyperplan optimal qui sépare les différentes classes dans un espace de caractéristiques de dimension potentiellement élevée. L'objectif fondamental est de déterminer un hyperplan qui maximise la marge, c'est-à-dire la distance minimale entre cet hyperplan et les points de données les plus proches de chaque classe, appelés vecteurs de support. [28]

Dans le cas où les deux classes sont linéairement séparables, comme illustré dans la figure, l'approche consiste à rechercher un hyperplan discriminant, représenté par l'équation suivante :

$$w * x + b = 0 \tag{2.1}$$

Les éléments de cette équation sont définis comme suit :

- w $(w1, w2, w3, \ldots, wn)$: le vecteur des poids .
- x $(x1, x2, x3, \ldots, xn)$: le vecteur des attributs .
- b : est le seuil du séparateur linéaire.

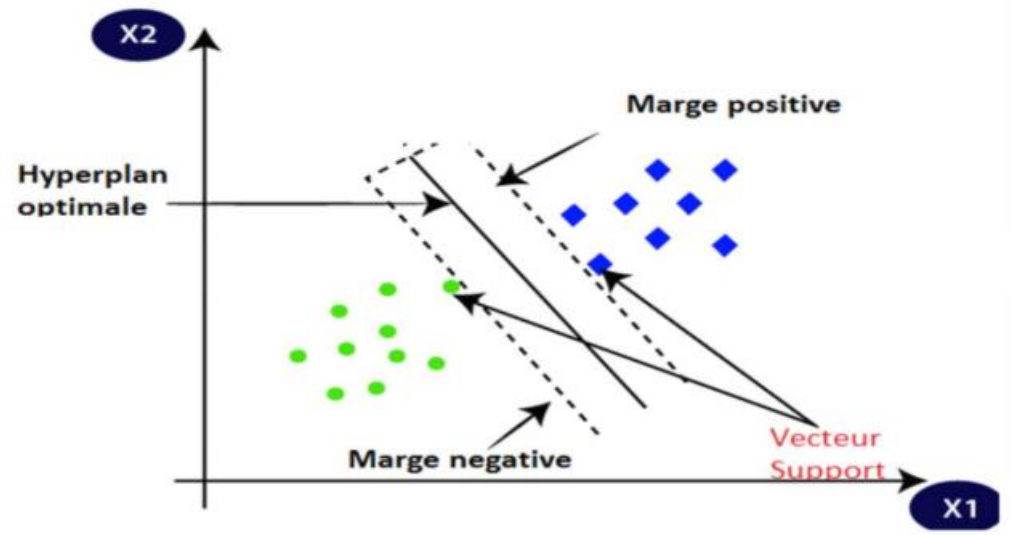

Figure 12- Exemple de classification par SVM

- **Hyperplan :** L'hyperplan divise les données en différentes classes. La tâche consiste à trouver l'hyperplan idéal qui minimise les erreurs de classification tout en maximise la marge entre les points de données des différentes classes.
- **Marge :** La distance entre l'hyperplan de séparation et les vecteurs de support les plus proches est représentée par la marge d'une SVM, et l'optimisation vise à trouver l'hyperplan qui maximise cette distance afin d'obtenir une meilleure séparation entre les classes et une meilleure généralisation des prédictions.
- **Vecteur support :** Les points de données les plus proches de l'hyperplan de séparation entre les différentes classes sont appelés vecteurs de support. La position et l'orientation de l'hyperplan optimal sont déterminées par ces vecteurs.

Linéarité et non linéarité

SVM linéaire

Un classificateur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire. Dans la suite, nous supposons que nos exemples sont donnés dans un format vectoriel. Notre espace d'entrée x est composé de n composantes. Si les données sont linéairement séparables, alors il existe un hyperplan d'équation [29].

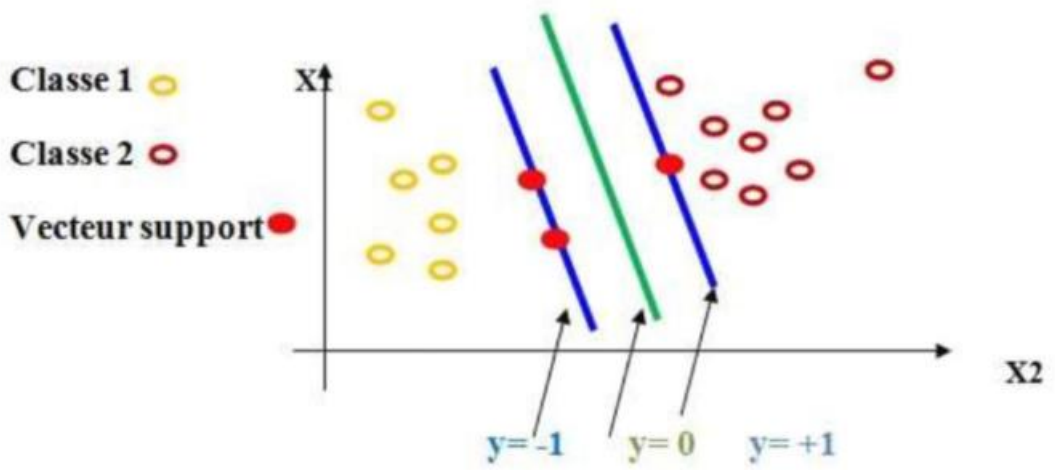

Figure 13 - Exemple graphique de données linéairement séparables

SVM non-linéaire

Un SVM non linéaire utilise des techniques pour traiter des données qui ne peuvent pas être séparées de manière linéaire dans l'espace d'origine. Pour cela, il projette les données dans un espace de dimension supérieure où elles peuvent potentiellement devenir linéairement séparables. Ainsi, le concept de séparation linéaire peut être appliqué dans cet espace de dimension supérieure. Formellement, dans le cas d'un SVM non linéaire utilisant une fonction noyau K, l'hyperplan séparateur s'exprime par la relation de la fonction de décision f(x) pour un nouvel échantillon x qui s'écrit de la forme suivante :

$$f(x) = \sum_{i=1}^{n} \propto_i y_i K(x_i, x) + b \qquad (2.2)$$

Où :

- n : représente le nombre total des échantillons d'apprentissage.
- $\propto i$: les coefficients optimisés par SVM (issues de la résolution du problème dual).
- y_i est l'étiquette de la classe de l'échantillon d'entraînement xi (yi= ± 1 pour les classes binaires).
- $K\ (x_i\ , x)$ Fonction noyau mesurant la similarité entre les échantillons xi et x dans un espace de features transformé.
- b est le biais.

Vapnik [30] a établi que toute fonction vérifiant les conditions de Mercer (conditions d'admissibilité) peut servir de noyau.

Les noyaux couramment employés en classification sont : le noyau linéaire, le noyau polynomial et le noyau Gaussien.

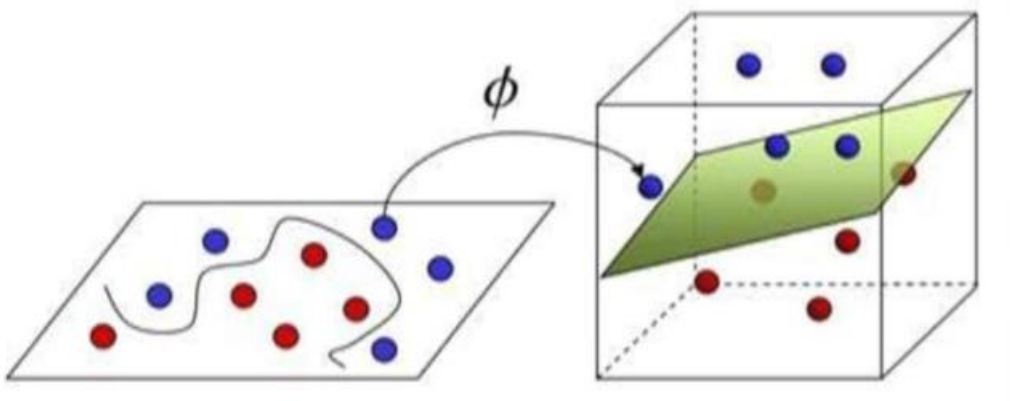

Figure 14 - Exemple de transformation des données non linéaires dans un espace de dimension

4. Forêts aléatoires et Arbre de décision

1. Arbre de décision

Un arbre de décision est un modèle prédictif hiérarchique où chaque nœud pose une question (ex. : âge < 30 ans ?), et chaque branche correspond à une réponse (oui/non). Ce processus se répète jusqu'à atteindre une feuille, qui donne la prédiction finale (classe ou valeur) [31].

Critères de séparation

Pour choisir la meilleure division à chaque nœud, on utilise des métriques mesurant la pureté des sous-ensembles :

Indice de Gini et Entropie

$$Gini = 1 - \sum_{k=1}^{m} f_k^2 \quad (2.3)$$

$$\text{Entropie} = -\sum_{k=1}^{m} f_k * \log_2(f_k) \quad (2.4)$$

- m : Nombre de classes possibles.
- f_k : proportion d'exemples appartenant à la classe dans le nœud [32].

Construction de l'arbre

L'algorithme CART (Classification And Regression Trees) est couramment utilisé. À chaque étape :

- Tester toutes les variables et divisions possibles,
- Choisir la division maximisant la pureté (via Gini ou gain),
- Répéter récursivement jusqu'à un critère d'arrêt : profondeur max, nombre minimal d'échantillons, gain trop faible, etc. [33].

Exemple d'un arbre de décision

La question posée est : Peut-on jouer un match de tennis aujourd'hui ?

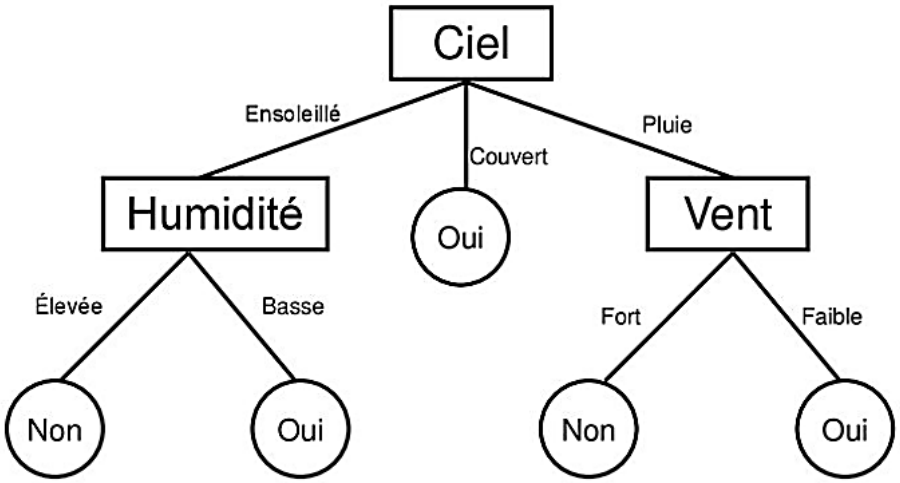

Figure 15 - exemple d'un arbre de décision

2. Forêt aléatoire (Random Forest)

Une forêt aléatoire est un ensemble d'arbres de décision construits de manière aléatoire :

- **Bagging** : on crée plusieurs sous-échantillons (bootstrap samples) des données.
- Sur chaque échantillon, on construit un arbre. Pour chaque nœud, on ne considère qu'un sous-ensemble aléatoire de variables (paramètre mtry) [34].
- Pour la prédiction finale, chaque arbre vote, et la majorité l'emporte (ou la moyenne pour la régression) [34].

Hyperparamètre mtry

- $mtry$ = nombre de variables candidates sélectionnées aléatoirement aux nœuds d'un arbre.
- Plus $mtry$ est petit → plus les arbres sont différents → meilleure robustesse [34].

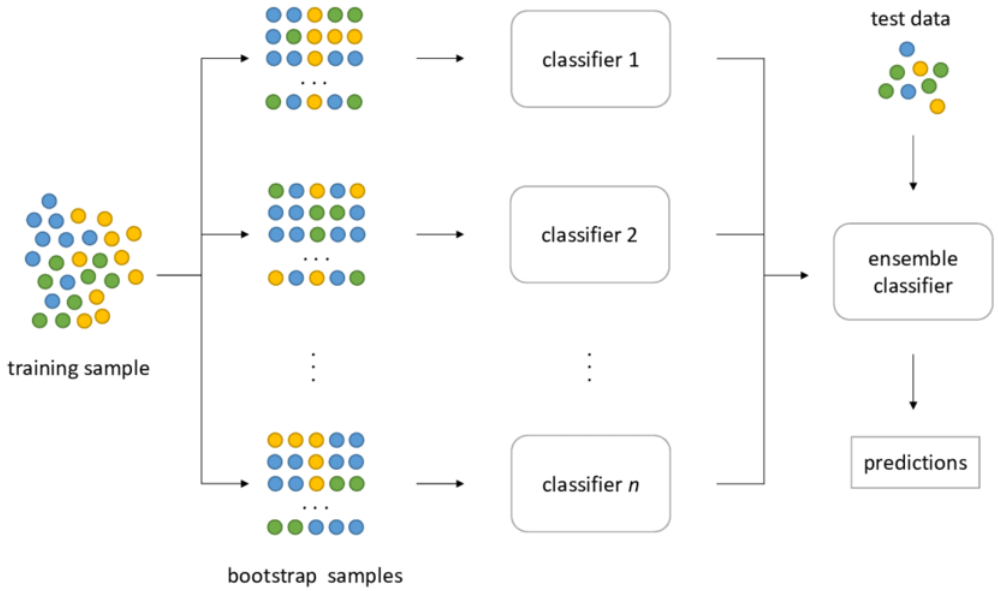

Figure 16 Processus de construction d'un Random Forest

5. Conclusion

Ce deuxième chapitre a été consacré à la présentation des méthodes de machine learning, en particulier les principes d'apprentissage automatique et de classification. Nous nous sommes focalisés sur les deux algorithmes testés et utilisés dans notre étude : les SVM (Support Vector Machines) et les forêts aléatoires (Random Forests) basées sur les arbres de décision. Ces modèles, choisis pour leur efficacité en classification supervisée, serviront de fondement au chapitre suivant, qui portera sur leur implémentation pratique pour la classification des cellules mammaires en malignes ou bénignes, dans le cadre du diagnostic assisté du cancer du sein.

Chapitre 3 Étude expérimentale des modèles de classification

1. Introduction

Après avoir introduit dans le chapitre précédent les fondements théoriques des algorithmes de classification, ce chapitre se concentre sur leur mise en pratique à travers une étude expérimentale. Il s'agit ici d'appliquer concrètement les méthodes étudiées sur un jeu de données médicales, afin d'en évaluer les performances et la pertinence dans un contexte réel.

2. Base de données utilisée

1. Présentation du dataset

Le jeu de données *Wisconsin Breast Cancer (Original)* provient des dossiers médicaux du Dr William H. Wolberg, du Centre hospitalier de l'Université du Wisconsin à Madison, et a été constitué entre 1989 et 1991 à partir d'analyses cytologiques réalisées par biopsie de nodules mammaires suspects [35]. Cette méthode, peu invasive et peu coûteuse, permet de prélever des cellules tumorales pour un examen microscopique. Chaque échantillon a été classé par un pathologiste comme bénin ou malin.

2. Contexte clinique

Les cellules prélevées sont notées sur une échelle de 1 à 10 selon **neuf critères cytologiques** fortement corrélés à la malignité :

- **Épaisseur amas** (*Clump Thickness*) : mesure l'épaisseur des amas cellulaires.
- **Taille cellule** (*Cell Size*) : évalue la dimension moyenne des cellules.
- **Forme cellule** (*Cell Shape*) : caractérise la régularité de la forme des cellules.
- **Adhésion** (*Marginal Adhesion*) : indique la capacité des cellules à adhérer entre elles.
- **Taille épithéliale** (*Single Epithelial Cell Size*) : mesure la taille des cellules épithéliales individuelles.
- **Noyaux nus** (*Bare Nuclei*) : dénombre les noyaux visibles sans cytoplasme.
- **Chromatie** (*Bland Chromatin*) : décrit l'apparence de la chromatine dans le noyau.
- **Nucléoles** (*Normal Nucleoli*) : évalue la dimension et la visibilité des nucléoles.
- **Mitose** (*Mitoses*) : compte le nombre de divisions cellulaires observées.

3. Extrait de données

Afin d'obtenir un aperçu global et aléatoire des données, nous avons extrait un échantillon de 10 lignes de la base de données :

Épaisseur amas	Taille cellule	Forme cellule	Adhésion	Taille épithéliale	Noyaux nus	Chromatie	Nucléoles	Mitose	Classe
2	1	1	1	3	1	2	1	1	2
4	10	4	7	3	10	9	10	1	4
4	1	1	1	2	1	3	2	1	2
4	3	3	1	2	1	3	3	1	2
5	1	1	1	2	2	3	3	1	2
5	1	1	1	2	1	1	1	1	2
1	1	1	1	2	1	3	1	1	2
7	1	2	3	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
4	4	4	2	2	3	2	1	1	2

Tableau 1 - Echantillon de 10 lignes de la base de données

3. Préparation des données

Avant toute phase d'analyse ou de modélisation, un nettoyage rigoureux du jeu de données s'impose afin de garantir la qualité et la fiabilité des résultats. Les étapes suivantes ont été appliquées au dataset.

1. Traitement des valeurs manquantes

Lors de l'importation, certaines cellules présentaient le caractère « ? », signalant l'absence de données. Ces valeurs ont été automatiquement interprétées comme manquantes. Pour éviter toute distorsion dans l'analyse, l'ensemble des lignes contenant au moins une valeur manquante a été supprimé.

2. Conversion des types de données

Toutes les colonnes numériques ont été converties au format entier (int) afin d'assurer une lecture correcte par les algorithmes de machine learning. Cette conversion permet d'effectuer des opérations mathématiques sans erreurs de typage.

3. Reformatage des étiquettes de classe

Dans le jeu de données original, la variable cible `class` prenait deux valeurs :

- `2` pour les tumeurs bénignes,
- `4` pour les tumeurs malignes.

Pour une meilleure lisibilité et une compatibilité accrue avec les modèles de classification, ces valeurs ont été renommées :

- `2` → `0` (tumeur bénigne),
- `4` → `1` (tumeur maligne).

4. Suppression des doublons

Une détection des doublons a été effectuée. Les lignes identiques sur l'ensemble des colonnes ont été supprimées afin d'éviter la redondance et de préserver l'intégrité statistique de l'échantillon.

5. Suppression de la colonne d'identifiant

La colonne `ID`, correspondant à un identifiant unique pour chaque observation, a été supprimée. Elle ne contient aucune information utile pour la classification et peut introduire du bruit dans les modèles prédictifs.

6. Définition de la variable cible

La colonne `class` a été définie comme la variable cible (output) du modèle de classification.

4. Analyse exploratoire des données

1. Analyse statistique

Le tableau suivant présente les statistiques descriptives (moyenne, médiane, écart-type, minimum et maximum) des principales caractéristiques extraites du jeu de données.

	Moyenne	Écart-type	Médiane	Minimum	Maximum
Épaisseur amas	5,378619	2,869029	5	1	10
Taille cellule	4,222717	3,25128	3	1	10
Forme cellule	4,273942	3,141494	3	1	10
Adhésion	3,746102	3,158413	3	1	10
Taille épithéliale	3,879733	2,456544	3	1	10
Noyaux nus	4,806236	3,880509	3	1	10
Chromatie	4,200445	2,651634	3	1	10
Nucléoles	3,828508	3,387146	2	1	10
Mitose	1,91314	2,068909	1	1	10

Tableau 2 Tableau des statistiques descriptives

Observations Globales

Dans la plupart des cas, la médiane est inférieure à la moyenne, indiquant une distribution asymétrique vers la droite. Par exemple, *Noyaux nus* a une moyenne de 4,81 mais une médiane de 3, ce qui suggère la présence de quelques valeurs très élevées.

Les écarts-types sont généralement élevés (souvent > 2), traduisant une forte variabilité entre les tumeurs. *Taille cellule* (écart-type de 3,25) et *Noyaux nus* (3,88) en sont de bons exemples.

Enfin, bien que toutes les variables varient de 1 à 10, leur répartition diffère. *Mitose* a une médiane de 1 mais un maximum de 10, ce qui montre que la majorité des cas sont faibles, avec quelques exceptions extrêmes.

2. Analyse des distributions des caractéristiques

Une série d'histogrammes a été générée, qui permet de comparer la répartition des valeurs pour chaque paramètre, en distinguant les deux classes de diagnostic, Chaque histogramme représente la fréquence d'occurrence des valeurs d'une caractéristique donnée.

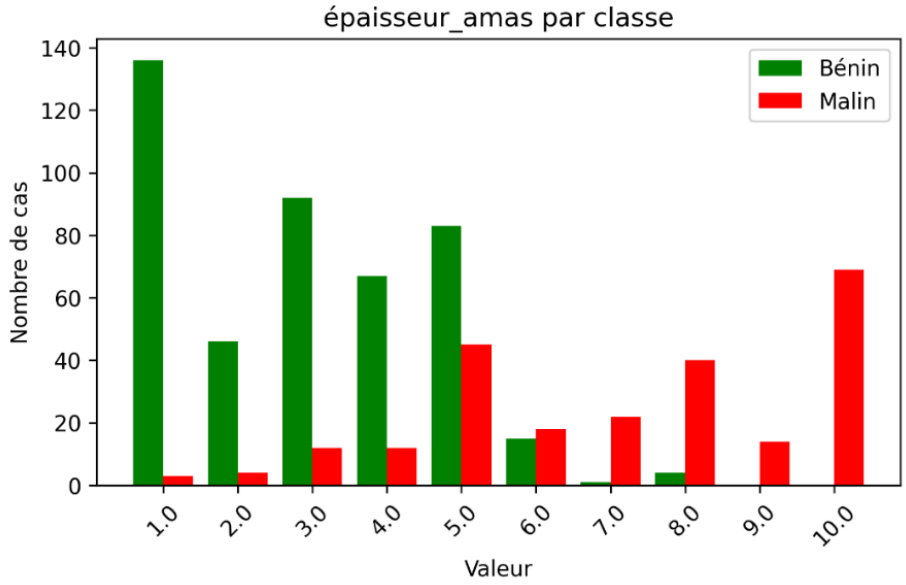

Figure 17 - Histogramme de l'épaisseur de l'amas

Les tumeurs bénignes présentent surtout des scores d'épaisseur 1–3 (rarement > 5), tandis que les tumeurs malignes se situent majoritairement entre 8 et 10, les valeurs ≥ 5 étant donc caractéristiques des formes malignes.

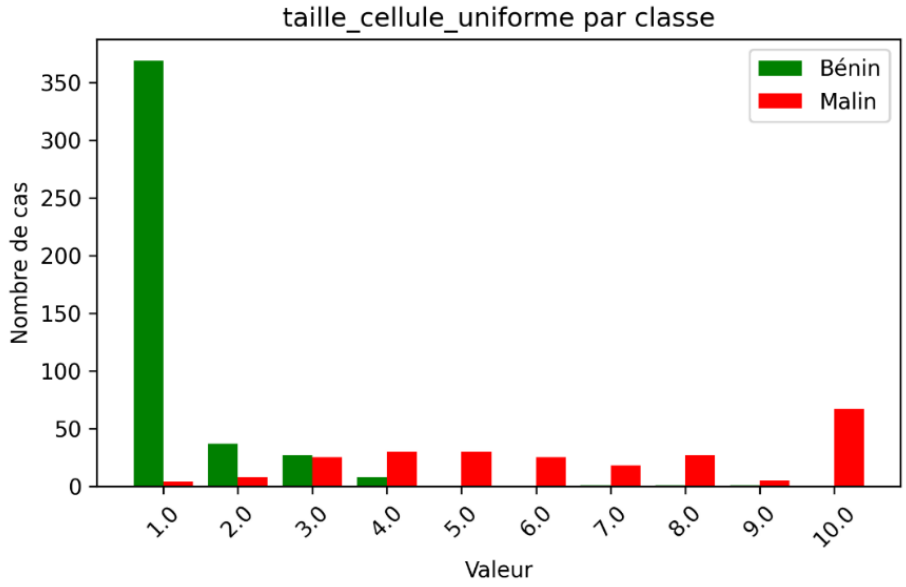

Figure 18 - Histogramme de la taille des cellules

Les cellules bénignes sont presque toutes de taille uniforme (score 1), alors que les malignes présentent une large variabilité avec des scores répartis de 4 à 10.

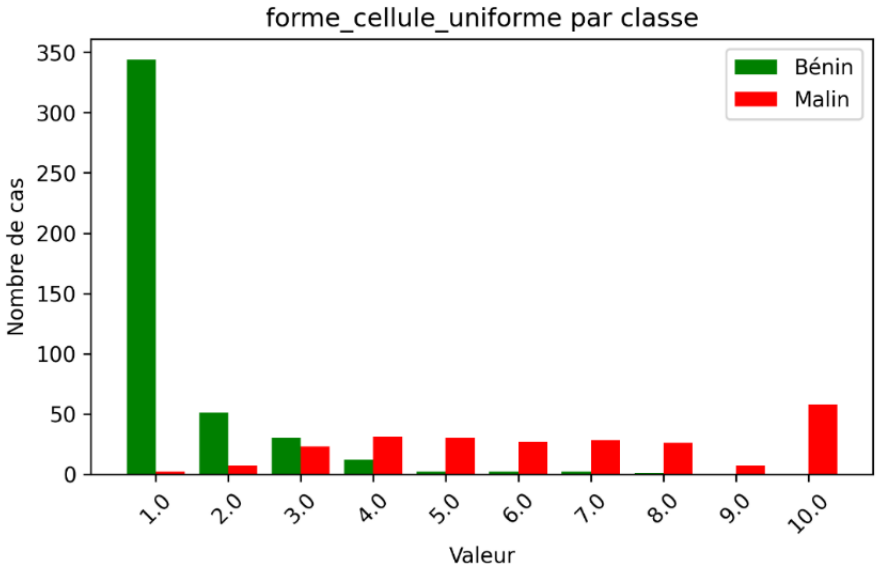

Figure 19 - Histogramme de la forme des cellules

Les cellules bénignes sont quasi toutes régulières (score 1, rarement > 2), tandis que les malignes, souvent asymétriques, culminent entre 6 et 8.

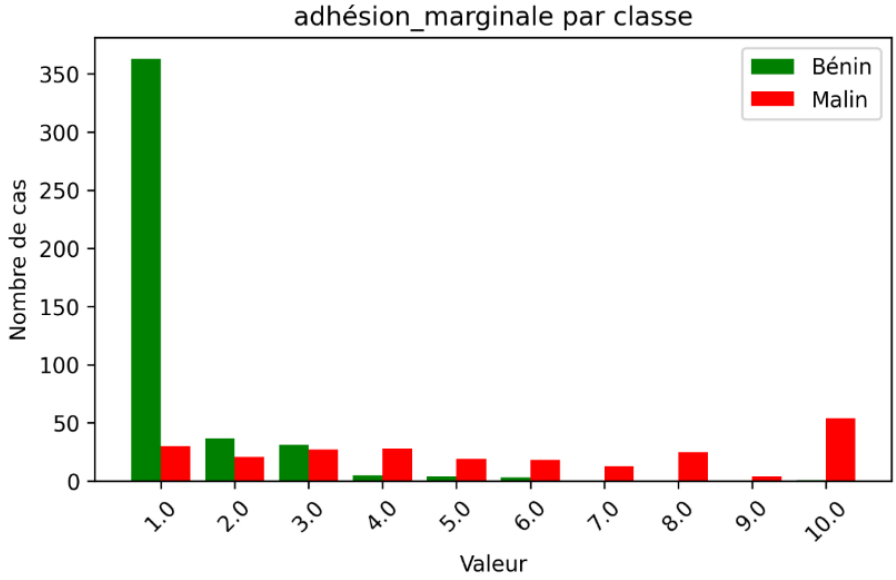

Figure 20 - Histogramme du degré d'adhésion cellulaire

Les cellules bénignes montrent principalement une adhésion faible (score 1, >3 rares), tandis que les malignes présentent une adhésion plus faible et variable (score 3–10).

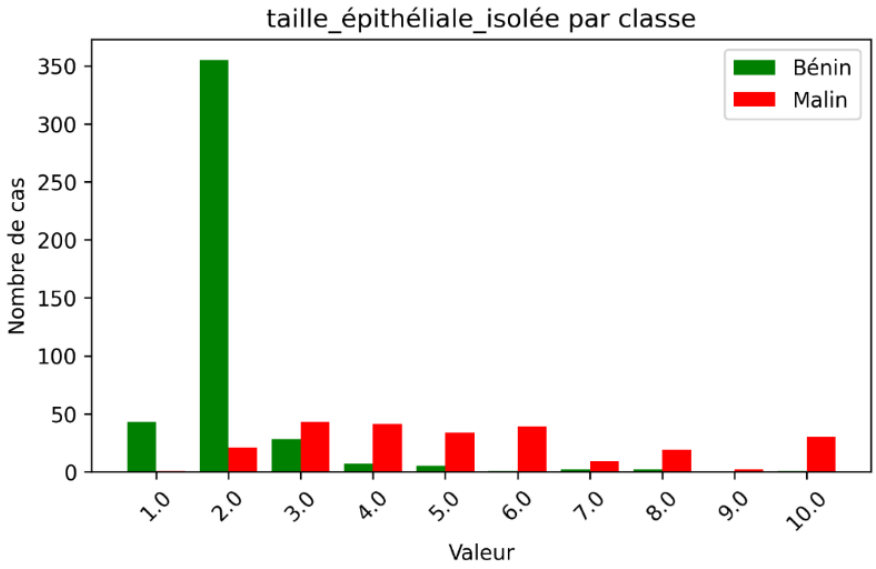

Figure 21 - Histogramme de la taille de l'épithéliale isolée

Les cellules bénignes ont une taille médiane de 2 avec peu de valeurs >3, tandis que les malignes présentent une taille médiane de 5.

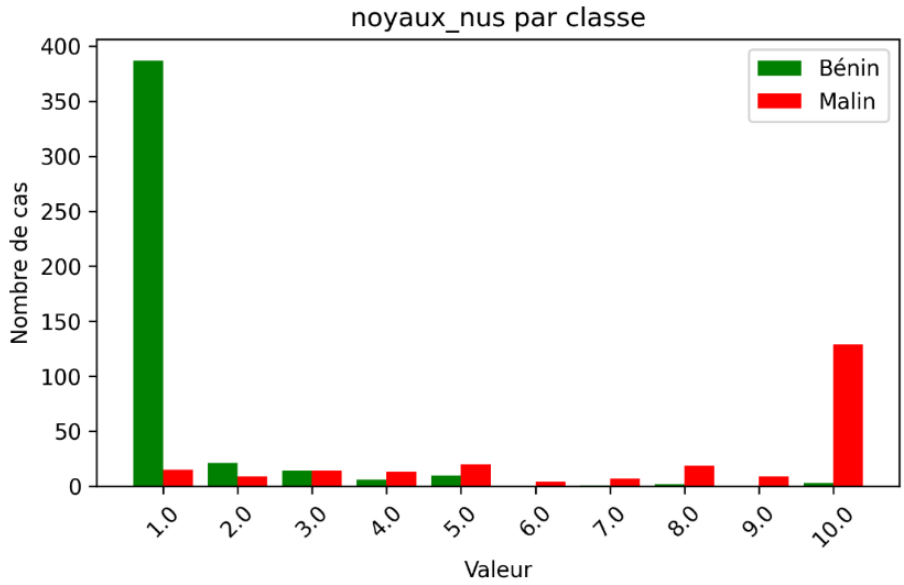

Figure 22 - Histogramme de la présence de noyaux nus

Les noyaux nus présentent un pic à 1 (90 %) pour les bénins avec des valeurs > 2 exceptionnelles, tandis que les malins montrent un pic à 10 (50 %).

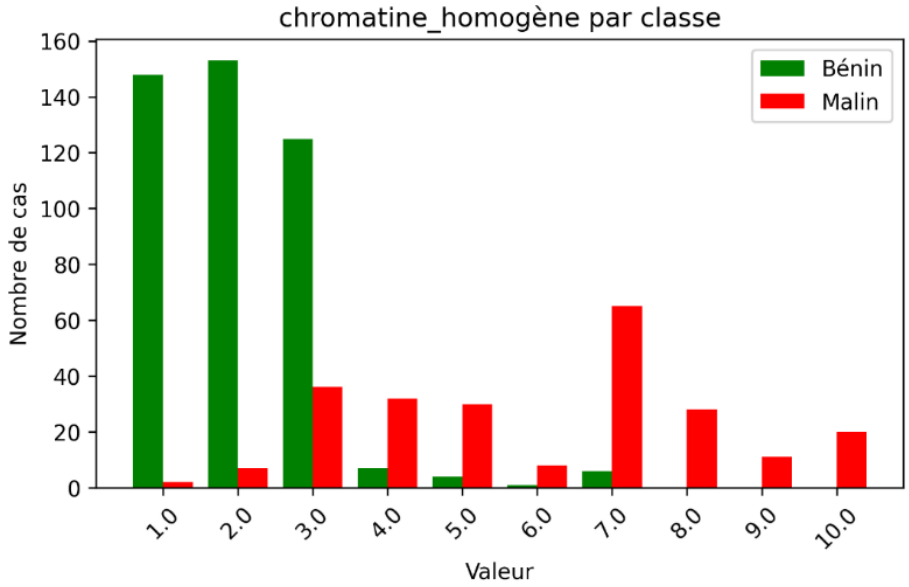

Figure 23 - Histogramme de la chromatine nucléaire

Chez les bénins, la chromatine présente un pic marqué à 1–3 (la plupart des cas), tandis que chez les malins, la distribution est centrée autour de 7, avec une médiane également à 7, indiquant une chromatine plus dense.

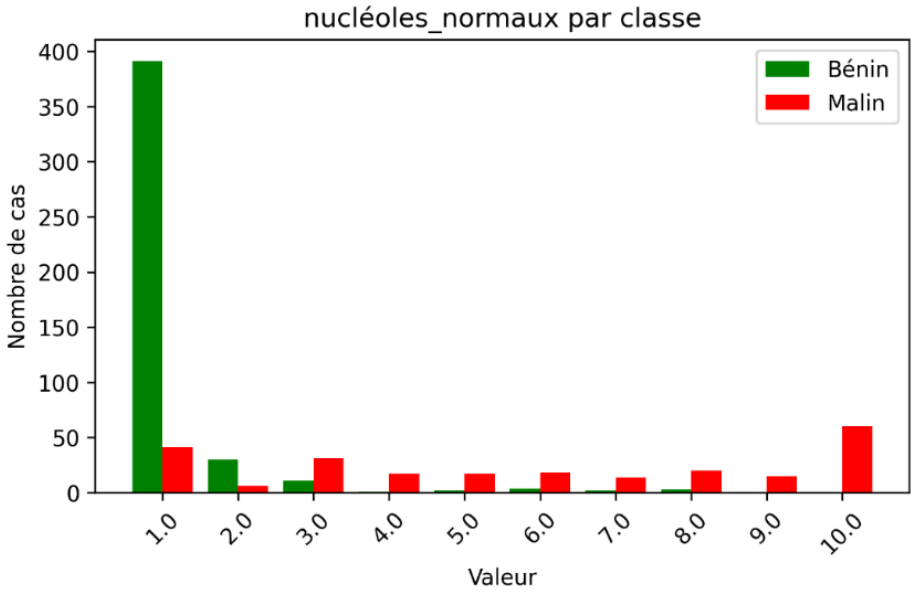

Figure 24 - Histogramme de la taille des nucléoles

Les tumeurs bénignes montrent un pic à 1 (80 % des cas), alors que les malignes ont une distribution plus étalée entre 3 et 10, avec un pic à 6, suggérant une fréquence plus élevée de nucléoles visibles.

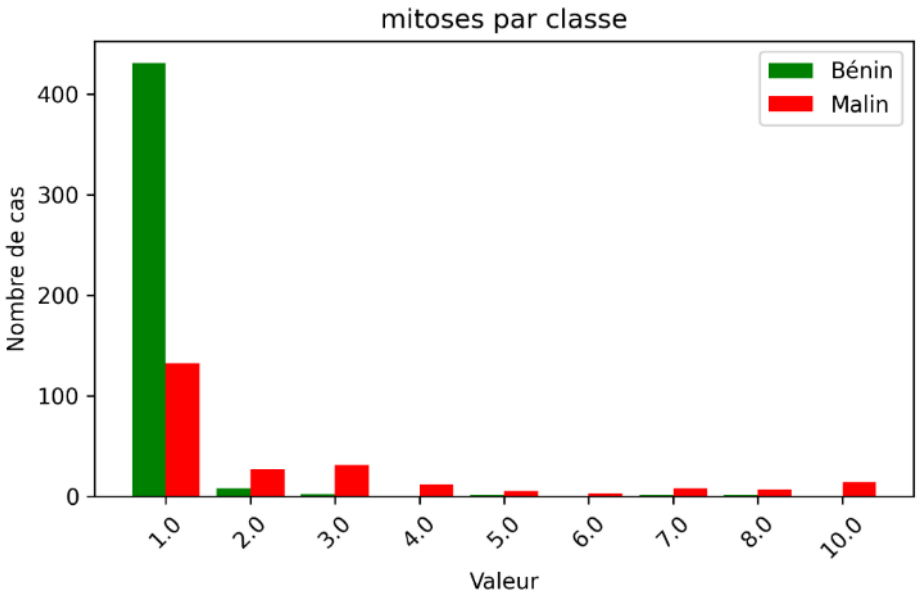

Figure 25 - Histogramme de l'activité mitotique

Les mitoses sont majoritairement absentes ou rares dans les bénins, avec un pic à 1 (95 %), tandis que chez les malins, bien que 50 % des cas présentent un score de 1, la distribution s'étend jusqu'à 10, reflétant une variabilité importante.

3. Analyse des Moyennes

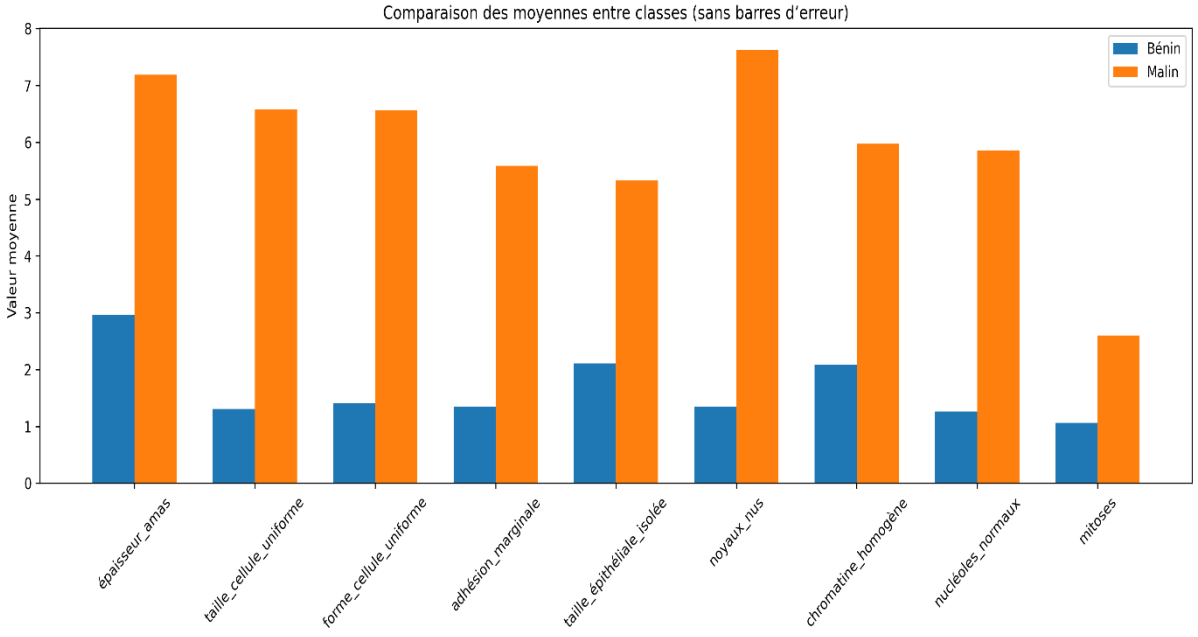

Figure 26- Moyennes entre classes

Le bar plot présente la comparaison des moyennes des 9 caractéristiques cellulaires entre tumeurs bénignes (B) et malignes (M). On observe que pour toutes les caractéristiques, les moyennes des tumeurs

malignes sont nettement supérieures à celles des bénignes, avec un écart minimal de +1.53 (Mitoses) et un maximum de +6.28 (Noyaux nus). Les différences les plus marquées concernent les noyaux nus, la taille et la forme cellulaire uniformes, ainsi que l'épaisseur des amas, indiquant leur fort pouvoir discriminant. En revanche, la caractéristique des mitoses montre une différence faible, suggérant une variabilité plus importante et un chevauchement possible entre les groupes. Les tumeurs bénignes présentent des valeurs basses et homogènes, traduisant leur uniformité morphologique. Ce graphique illustre clairement la séparation entre B et M sur la majorité des critères, mais il convient de compléter cette analyse par des mesures de dispersion et des représentations des distributions pour éviter des interprétations biaisées.

Interprétation médicale

Les noyaux nus, avec un score moyen très élevé dans les tumeurs malignes (+6.28), traduisent une instabilité génétique caractéristique des cancers agressifs. La taille et la forme cellulaires uniformes, également fortement augmentées, reflètent une hétérogénéité morphologique typique des cellules tumorales malignes peu différenciées, souvent associée à un pronostic défavorable. La faible différence observée pour les mitoses suggère que ce marqueur seul est insuffisant pour distinguer clairement les tumeurs bénignes des malignes, nécessitant une approche combinée avec d'autres critères. La forte homogénéité des scores dans les tumeurs bénignes confirme leur nature stable et bien différenciée.

4. Analyse des Corrélations des Paramètres

Les heatmaps permettent de visualiser les relations linéaires entre les différentes caractéristiques mesurées dans les tumeurs bénignes et malignes. Chaque case représente le coefficient de corrélation de Pearson entre deux variables, avec une échelle allant de -1 (corrélation négative parfaite) à +1 (corrélation positive parfaite).

Nous avons représenté ces corrélations à travers des heatmaps distinctes, chacune illustrant les relations internes des classes bénignes et malignes, ainsi que les corrélations croisées entre les paramètres des deux groupes.

Corrélation Bénin

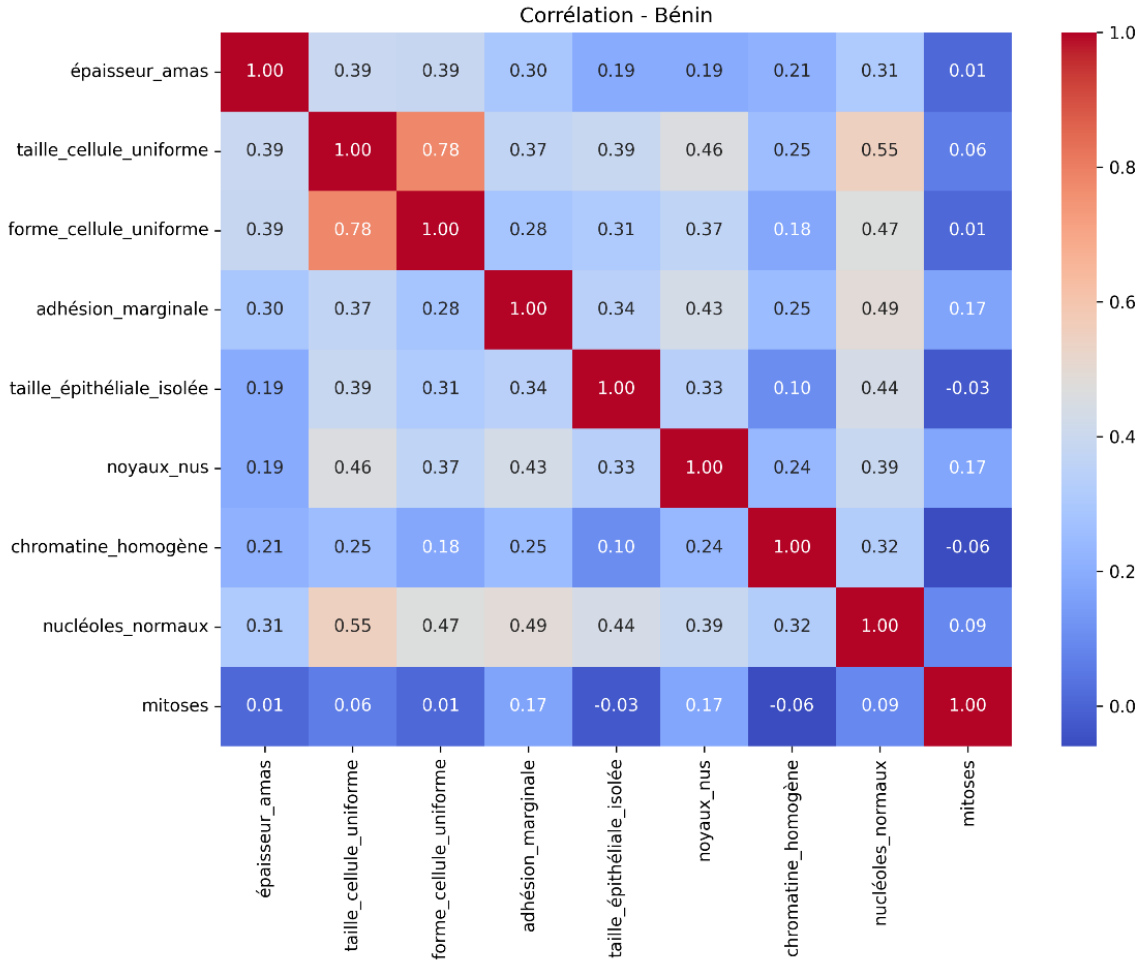

Figure 27 - Heatmaps corrélations bénignes

Corrélations fortes

Taille uniforme ↔ Forme uniforme (0.78)

Les cellules gardent une taille et une forme régulières. Cela montre une grande cohérence dans leur structure, ce qui est typique des tissus non cancéreux.

Corrélations modérées

Nucléoles ↔ Taille des cellules (0.55)

Plus les cellules sont grandes, plus les structures internes comme les nucléoles sont visibles et développées.

Adhésion marginale ↔ Noyaux nus (0.43)

Ces deux éléments semblent évoluer ensemble, indiquant une certaine cohésion cellulaire.

Chromatine ↔ Nucléoles (0.32)

Des liens existent entre la texture interne du noyau et les éléments qu'il contient, mais ils sont moins marqués.

Corrélations faibles

Mitoses (≤ 0.17)

Le nombre de divisions cellulaires ne suit pas la logique des autres paramètres. Cela montre que les cellules se divisent peu, ou de façon indépendante des autres caractéristiques.

Conclusion pour les cas bénins

Les paramètres évoluent de façon cohérente et prévisible. Cela reflète la stabilité biologique des tumeurs bénignes. Seules les mitoses semblent agir de façon isolée.

Corrélation Malin

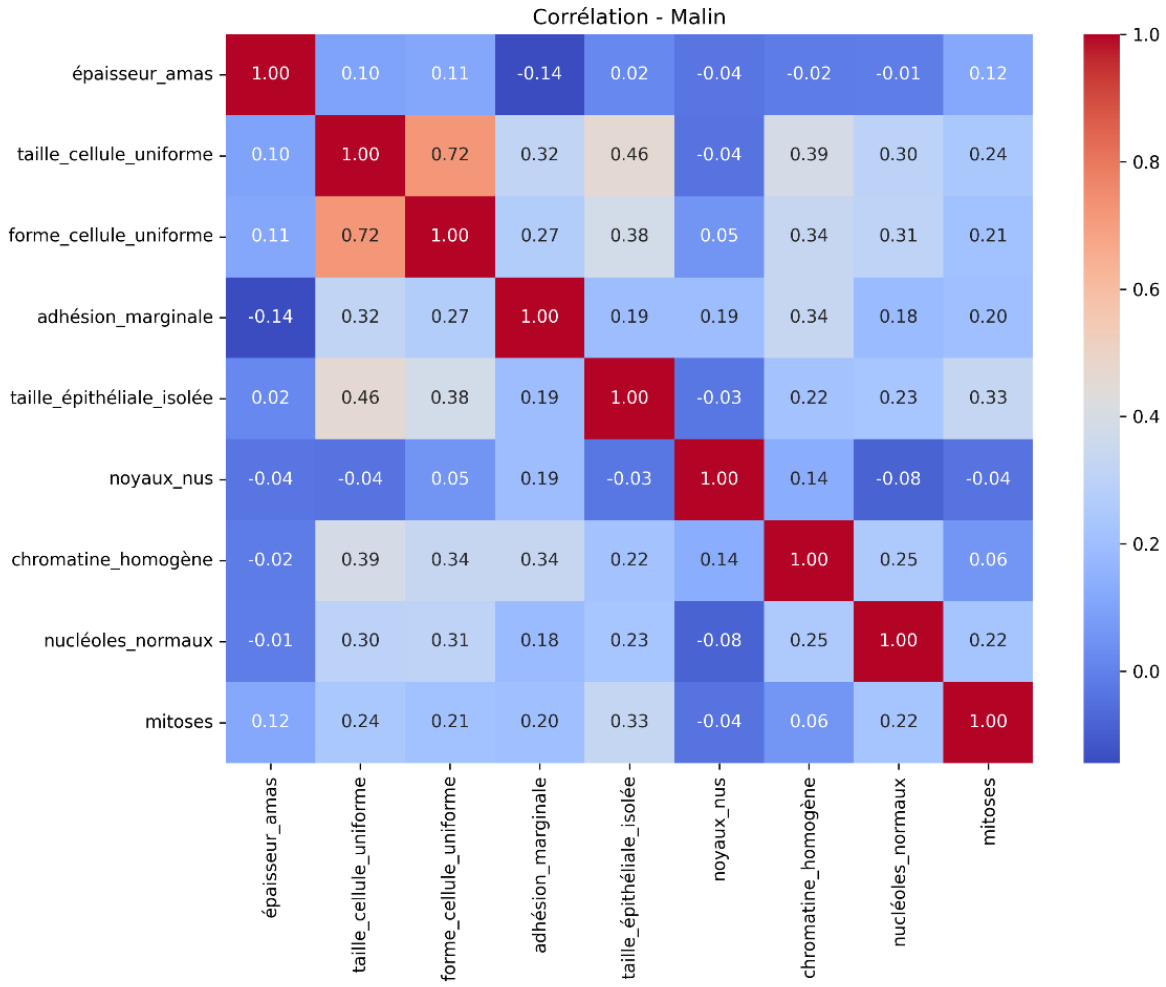

Figure 28 - Heatmaps corrélations malignes

Corrélations fortes

Taille uniforme ↔ Forme uniforme (0.72)

La relation reste présente, mais moins forte. Les cellules perdent leur régularité, ce qui reflète une structure plus chaotique.

Corrélations modérées

Taille uniforme ↔ Taille cellules isolées (0.46)

Cela montre que l'irrégularité des cellules affecte aussi l'environnement immédiat.

Forme uniforme ↔ Chromatine (0.34)

Les anomalies internes perturbent l'apparence extérieure des cellules.

Corrélations faibles ou absentes

Épaisseur amas (~0.1)

Elle évolue de façon indépendante, ce qui en fait un critère intéressant pour la différenciation.

Noyaux nus ↔ Taille uniforme (-0.04)

Aucun lien. Ces deux éléments ne réagissent pas ensemble, ce qui confirme la variabilité du comportement cellulaire.

Conclusion pour les cas malins

On observe moins de cohérence entre les paramètres, ce qui illustre la grande hétérogénéité des cellules cancéreuses. Certains indicateurs évoluent de manière autonome, et peuvent donc aider à mieux distinguer les tumeurs malignes.

Corrélation croisée

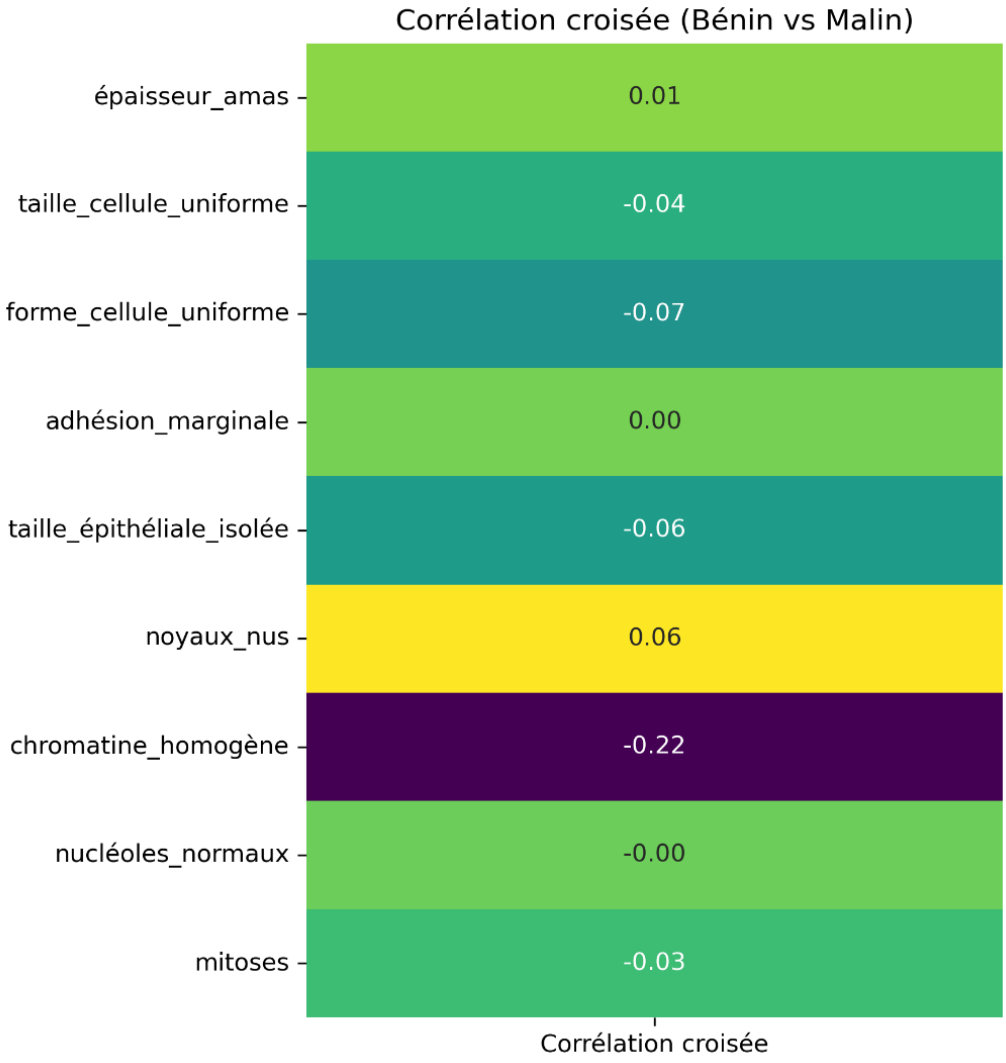

Figure 29 - Corrélation - croisé

Quand on compare les deux types de tumeurs, la majorité des corrélations sont :

- **Très faibles**
- **Souvent négatives**
- **Parfois quasi nulles**

Cela signifie que le comportement d'un paramètre dans une tumeur bénigne ne prédit rien du tout dans une tumeur maligne. Quelques exemples :

Chromatine homogène (-0.22)

Fortement organisée chez les bénins, complètement déréglée chez les malins.

Forme uniforme (-0.07)

La régularité des formes disparaît dans les tumeurs cancéreuses.

Épaisseur amas (0.01), Adhésion marginale (0.00), Mitoses (-0.03)

Ces valeurs montrent zéro lien entre les deux classes.

Anlayse

Les tumeurs bénignes montrent des relations logiques, stables et structurées.
Les tumeurs malignes, au contraire, présentent une forte variabilité et des relations beaucoup plus faibles ou inversées.

Cela suggère que ces deux types de tumeurs ne suivent pas le même mécanisme de transformation cellulaire et Il n'y a pas de progression linéaire.

5. Séparation des données

Pour l'expérimentation des modèles de classification, il est essentiel de diviser le jeu de données en plusieurs sous-ensembles. Nous avons choisi d'explorer trois types de séparation des données :

1. Division Entraînement / Test (80 % – 20 %)

Cette méthode consiste à utiliser 80 % des données pour l'apprentissage du modèle, et 20 % pour son évaluation finale. Elle est simple à mettre en œuvre et adaptée lorsque le volume de données est important.

2. Division Entraînement / Validation / Test (60% – 20% – 20%)

Cette approche répartit les données en trois sous-ensembles : 60 % pour l'entraînement, 20 % pour la validation (utilisée pour régler les paramètres du modèle), et 20 % pour le test final. Elle est recommandée dans les cas où une optimisation précise du modèle est nécessaire, tout en préservant un ensemble indépendant pour l'évaluation objective.

3. Division Entraînement / Validation / Test (70% – 10% – 20%)

Lorsque le jeu de données est restreint, il peut être avantageux d'allouer une plus grande proportion de données à l'apprentissage (70 %), tout en conservant une petite portion pour la validation (10 %) et une portion stable pour le test (20 %). Cette stratégie permet de maximiser l'efficacité de l'apprentissage tout en conservant une évaluation fiable.

6. Application des modèles de classification

Après avoir nettoyé et préparé les données, la prochaine étape consiste à appliquer les modèles de classification présentés dans le chapitre précédent, à savoir le SVM et la forêt aléatoire.

1. Outils d'évaluation

Afin de juger objectivement la performance des modèles de classification appliqués aux données, plusieurs métriques d'évaluation sont utilisées. Ces outils permettent de comparer les algorithmes entre eux, mais aussi de mieux comprendre leurs points forts et leurs limites dans le contexte du diagnostic du cancer du sein.

Matrice de confusion

La matrice de confusion permet de visualiser les résultats d'un modèle en termes de prédictions correctes et incorrectes, en distinguant :

- **Vrais positifs (VP)** : cas malins correctement identifiés,
- **Faux positifs (FP)** : cas bénins incorrectement identifiés comme malins,
- **Vrais négatifs (VN)** : cas bénins correctement identifiés,
- **Faux négatifs (FN)** : cas malins non détectés.

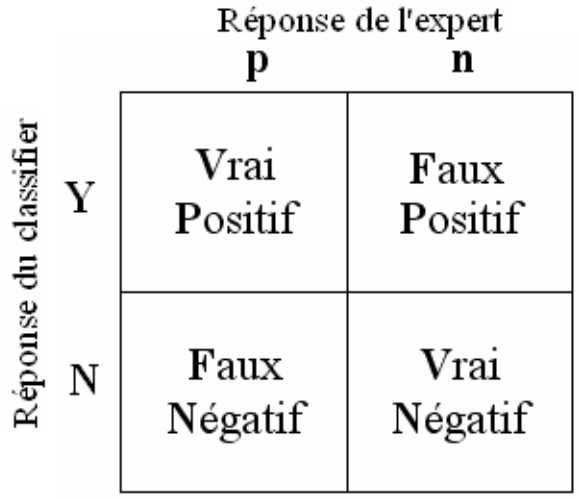

Figure 30 - Matrice de confusion

Précision (Precision)

$$\text{Précision} = \frac{VP}{VP + FP} \tag{3.1}$$

Mesure la proportion de cas réellement malins parmi ceux prédits comme tels. Une précision élevée signifie peu de fausses alertes.

Rappel (Recall ou Sensibilité)

$$Rappel = \frac{VP}{VP + FN} \tag{3.2}$$

Indique la capacité du modèle à détecter les cas malins. Un bon rappel limite les oublis de diagnostics.

Score F1

$$F1 = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel} \tag{3.3}$$

C'est une moyenne harmonique entre la précision et le rappel, utile lorsque les classes sont déséquilibrées.

Exactitude (Accuracy)

$$\text{Exactitude} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{3.4}$$

Mesure la proportion totale de prédictions correctes. Elle peut être trompeuse si une classe domine.

Aire sous la courbe ROC (AUC - ROC)

L'AUC évalue la capacité d'un modèle à distinguer les deux classes (bénin/malin), quelle que soit la probabilité de seuil choisie. Plus l'AUC est proche de 1, plus le modèle est performant.

2. Application du Modèle Random Forest

Random Forest – Division 80/20

Pour cette première expérimentation, le jeu de données a été divisé en deux sous-ensembles : 80 % pour l'entraînement et 20 % pour le test :

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.966	0.966	0.966	88
Maligne	0.936	0.936	0.936	47
Exactitude	0.956	0.956	0.956	—

Tableau 3 - Résultats de classification - Random Forest - Division 80/20

Matrice de confusion

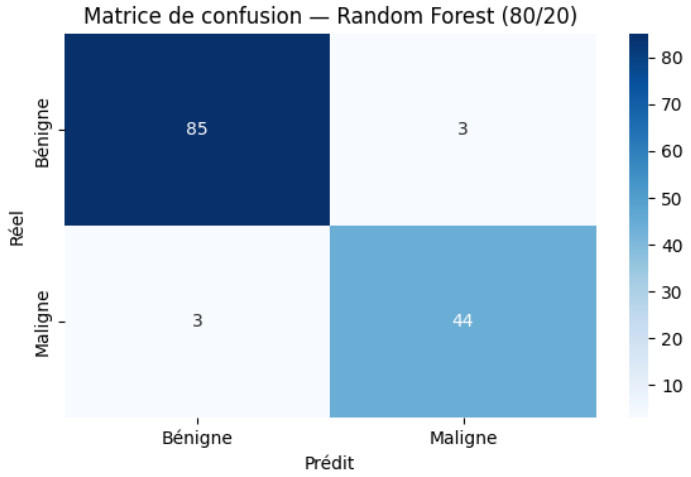

Figure 31 - Matrice de confusion - RF - 80/20

Courbe ROC

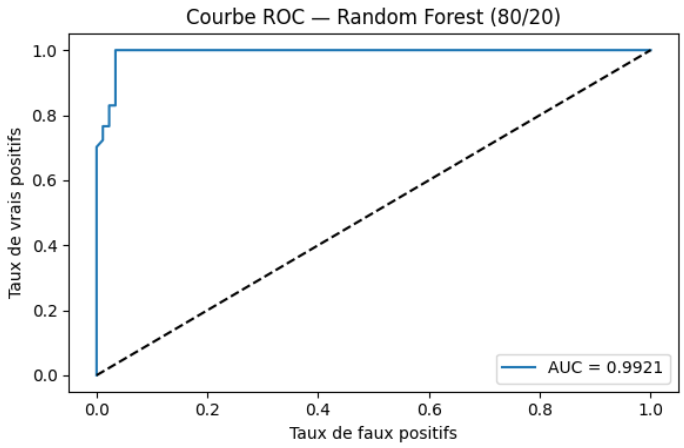

Figure 32 - Courbe ROC Random forest 80/20

Random Forest – Division 60/20/20

Dans cette configuration, le jeu de données est réparti en trois sous-ensembles: 60 % pour l'entraînement, 20 % pour la validation, et 20 % pour le test.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.966	0.966	0.966	88
Maligne	0.936	0.936	0.936	47
Exactitude	0.956	0.956	0.956	—

Tableau 4 - Résultats de classification Division 60/20/20

Matrice de confusion

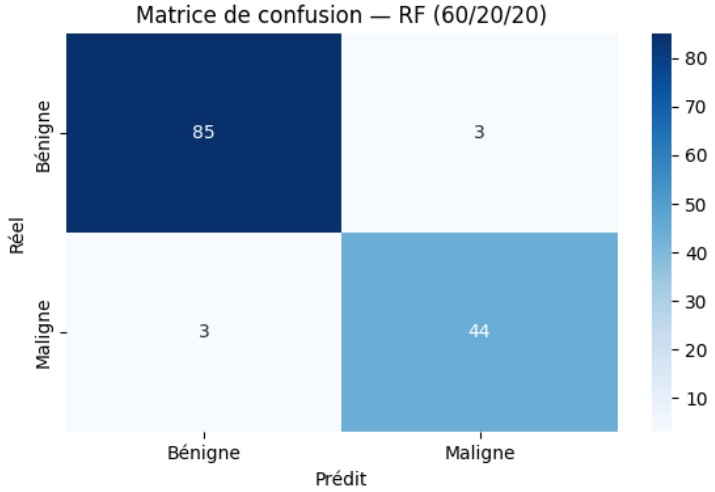

Figure 33 - Matrice de confusion - RF - 60/20/20

Courbe ROC

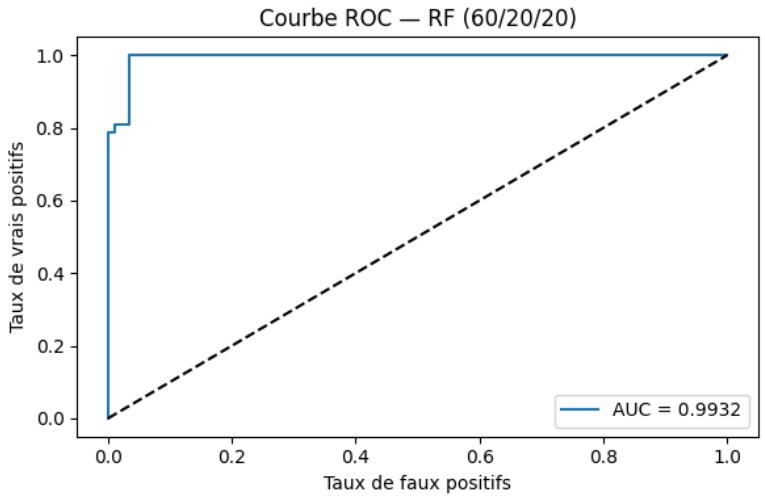

Figure 34 - Courbe ROC Ranfom forest 80/20

Random Forest – Division 70/10/20

Cette configuration est particulièrement adaptée aux petits jeux de données, car elle maximise les données disponibles pour l'apprentissage (70 %), tout en conservant une petite portion pour la validation (10 %) et une portion fixe pour le test (20 %).

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.966	0.966	0.966	88
Maligne	0.936	0.936	0.936	47
Exactitude	0.956	0.956	0.956	—

Tableau 5 - Résultats de classification Division 70/10/20

Matrice de confusion

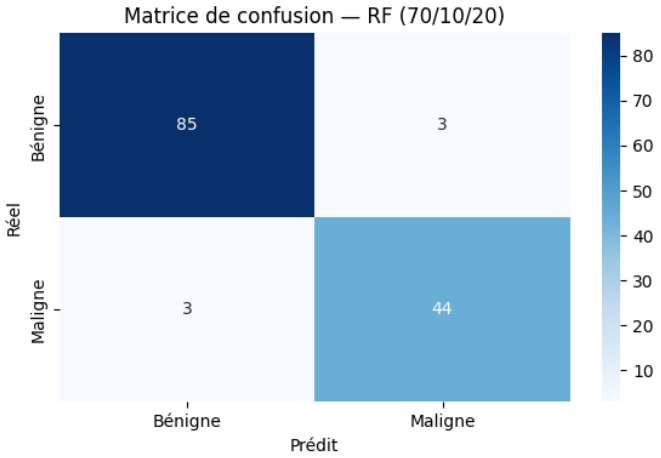

Figure 35 - Matrice de confusion - RF - 70/10/20

Courbe ROC

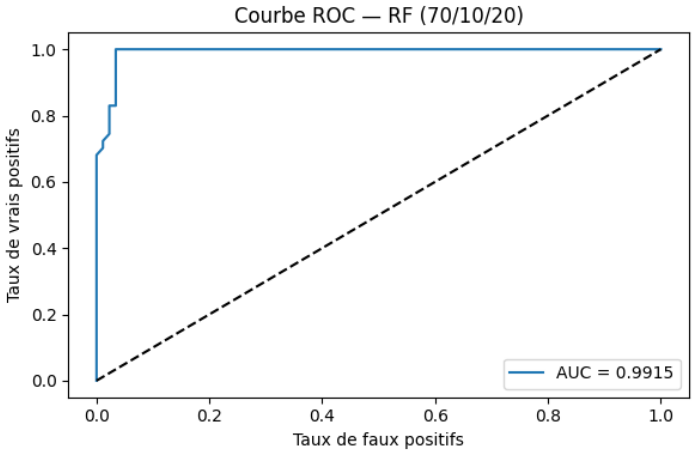

Figure 36 - Courbe ROC Ranfom forest 70/10/20

Analyse

Le modèle Random Forest affiche des performances constantes à travers les trois découpages de données, avec une exactitude de 95,6 %, un F1-score de 0.936 pour la classe maligne, et une AUC estimée à 0.96. La matrice de confusion montre cependant la présence de trois faux négatifs dans chaque cas, c'est-à-dire trois cas de tumeurs malignes non détectés, ce qui constitue un point critique dans un contexte médical. En effet, les faux négatifs représentent un risque élevé, car ils peuvent retarder la prise en charge d'un cancer réel. Malgré une bonne précision globale et une grande stabilité du modèle quel que soit le découpage, cette tendance à manquer certains cas malins limite la fiabilité du Random Forest dans des applications où la sensibilité diagnostique est essentielle.

3. Application du Modèle SVM

SVM – Division 80/20

Le modèle SVM (Support Vector Machine) a été entraîné sur 80 % des données et évalué sur 20 % restantes.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.977	0.966	0.971	88
Maligne	0.938	0.957	0.947	47
Exactitude	0.963	0.963	0.963	—

Tableau 6 - Résultats de classification SVM – Division 80/20

Matrice de confusion

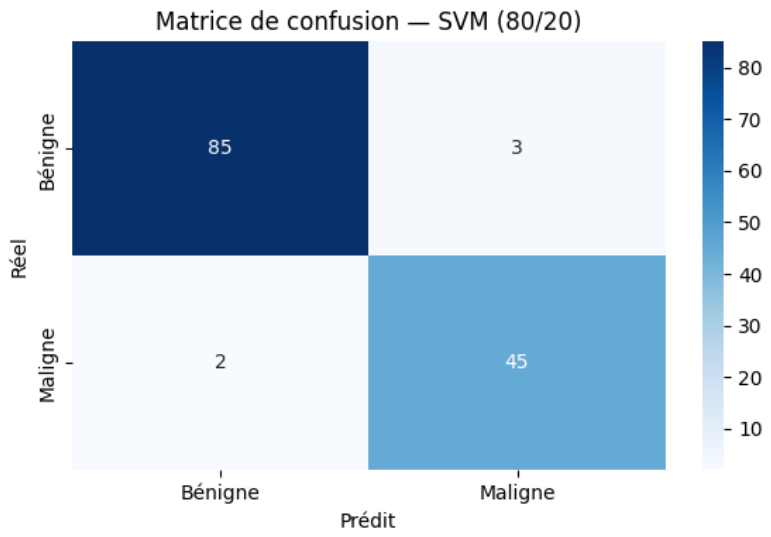

Figure 37 - Matrice de confusion - SVM - 80/20

Courbe ROC

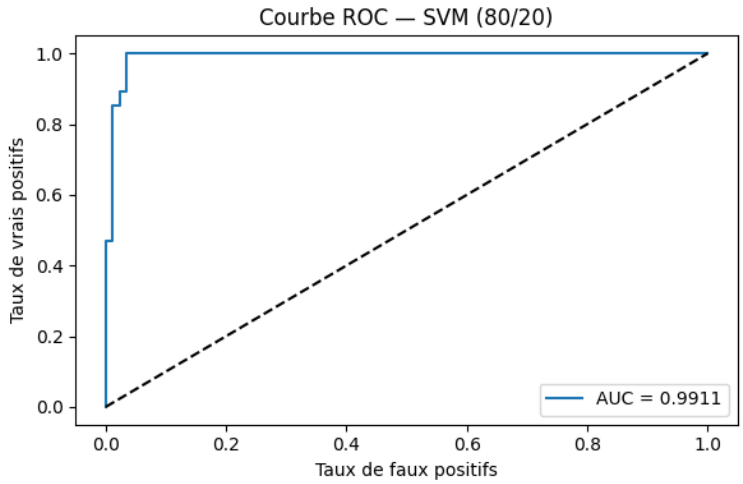

Figure 38 - Courbe ROC - SVM 80/20

SVM – Division 60/20/20

Avec cette stratégie, le jeu de données est séparé en 60 % pour l'entraînement, 20 % pour la validation, et 20 % pour le test.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.977	0.966	0.971	88
Maligne	0.938	0.957	0.947	47
Exactitude	0.963	0.963	0.963	—

Tableau 7 - Résultats de classification SVM – Division 60/20/20

Matrice de confusion

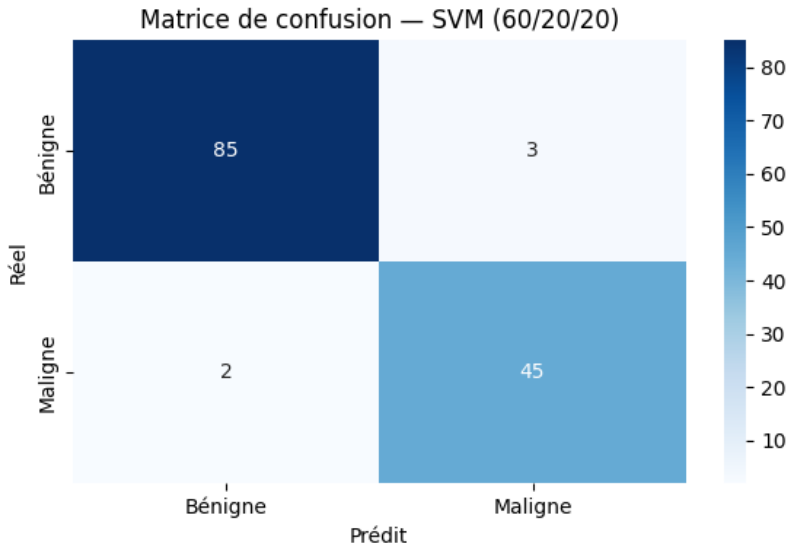

Figure 39 - Matrice de confusion - SVM - 60/20/20

Courbe ROC

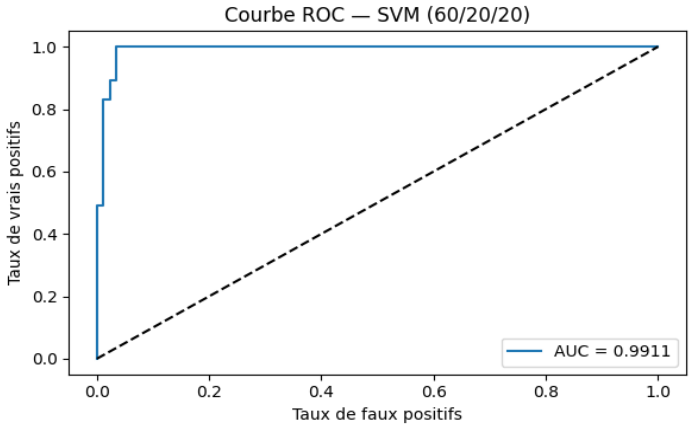

Figure 40 - Courbe ROC - SVM 60/20/20

SVM – Division 70/10/20

Cette configuration répartit les données en 70 % pour l'apprentissage, 10 % pour la validation, et 20 % pour le test.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.977	0.966	0.971	88
Maligne	0.938	0.957	0.947	47
Exactitude	0.963	0.963	0.963	—

Tableau 8 - Résultats de classification SVM – Division 70/10/20

Matrice de confusion

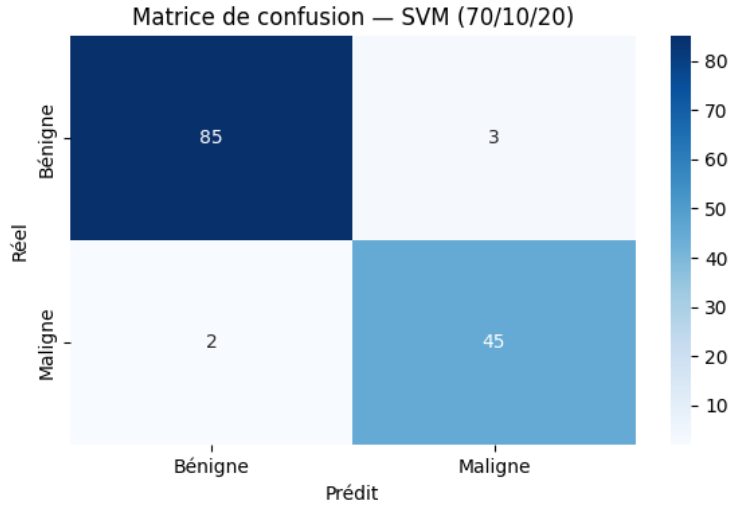

Figure 41 – Matrice de confusion – SVM – 70/10/20

Courbe ROC

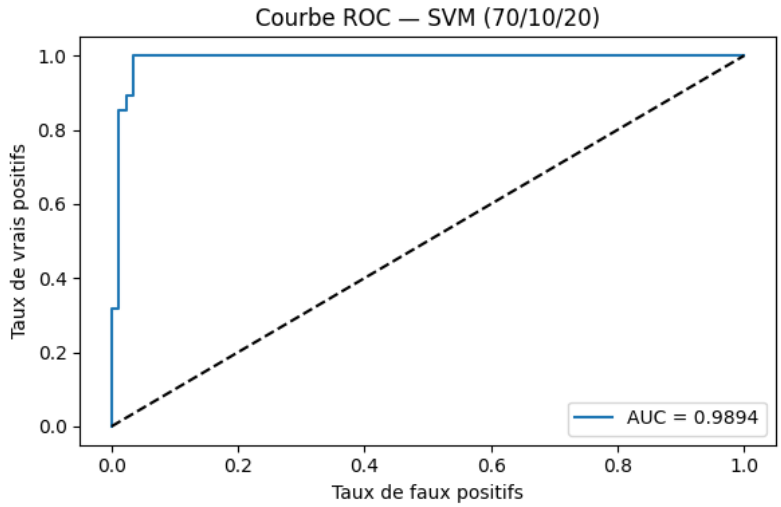

Figure 42 - Courbe ROC – SVM - 70/10/20

Analyse Générale

Le modèle SVM affiche des performances constantes à travers les trois découpages de données, avec une exactitude de 95,6 %, un F1-score de 0.936 pour la classe maligne, et une AUC estimée à 0.96. La matrice de confusion montre cependant la présence de trois faux négatifs dans chaque cas, c'est-à-dire trois cas de tumeurs malignes non détectés, ce qui constitue un point critique dans un contexte médical. En effet, les faux négatifs représentent un risque élevé, car ils peuvent retarder la prise en

charge d'un cancer réel. Malgré une bonne précision globale et une grande stabilité du modèle quel que soit le découpage, cette tendance à manquer certains cas malins limite la fiabilité du Random Forest dans des applications où la sensibilité diagnostique est essentielle.

7. Comparaison finale des modèles

En comparant les deux modèles, le SVM se distingue par une meilleure détection des cas malins, avec moins de faux négatifs (2 contre 3 pour Random Forest), un rappel plus élevé (0.957 contre 0.936), et une AUC supérieure (~0.98 contre ~0.96). Bien que Random Forest reste stable et performant, ces différences, bien que légères, sont cruciales en contexte médical où manquer un cas de cancer peut avoir des conséquences graves. Ainsi, le SVM s'impose comme le modèle le plus adapté pour un système d'aide au diagnostic fiable et sensible.

Le tableau ci-dessous résume les résultats obtenus

Division	Classe	Précision (RF)	Rappel (RF)	F1-score (RF)	Précision (SVM)	Rappel (SVM)	F1-score (SVM)	Support
80% Entraînement / 20% Test	Bénigne	0.966	0.966	0.966	0.977	0.966	0.971	88
	Maligne	0.936	0.936	0.936	0.938	0.957	0.947	47
	Exact.	0.956	0.956	0.956	0.963	0.963	0.963	—
60% / 20% / 20% (Train/Val/Test)	Bénigne	0.966	0.966	0.966	0.977	0.966	0.971	88
	Maligne	0.936	0.936	0.936	0.938	0.957	0.947	47
	Exact.	0.956	0.956	0.956	0.963	0.963	0.963	—
70% / 10% / 20% (Train/Val/Test)	Bénigne	0.966	0.966	0.966	0.977	0.966	0.971	88
	Maligne	0.936	0.936	0.936	0.938	0.957	0.947	47
	Exact.	0.956	0.956	0.956	0.963	0.963	0.963	—

Tableau 9 - Résumer des résultats obtenus

8. Évaluation de la performance en fonction du pourcentage de séparation des données

Le travail présenté dans cette section a pour objectif d'analyser l'impact de la proportion de séparation des données (split entre apprentissage et test) sur la performance des modèles SVM et Random Forest. En faisant varier systématiquement le pourcentage réservé au jeu de test, nous cherchons à identifier le seuil optimal qui permet aux deux modèles de maintenir une exactitude élevée tout en assurant une bonne capacité de généralisation. Cette étude comparative vise à localiser la meilleure configuration de division des données, afin d'optimiser l'entraînement tout en minimisant les erreurs de prédiction, notamment les faux négatifs, critiques dans un contexte de diagnostic médical.

1. Variation de l'exactitude du SVM selon la proportion du jeu de test

Le graphique ci-dessous illustre l'évolution de l'exactitude (accuracy) du modèle SVM en fonction du pourcentage de données alloué au test, allant de 10 % à 90 %.

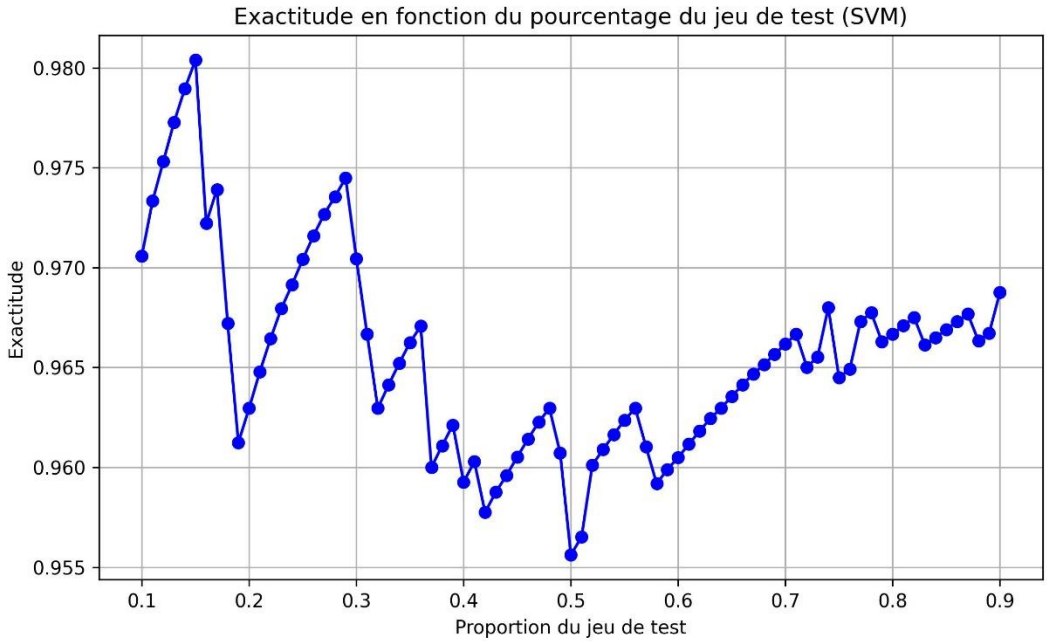

Figure 43 - Exactitude en focntion du pourcentage de tes

D'après le graphique ci-dessus, la meilleure performance a été obtenue avec une proportion de 15 % du jeu de données dédiée au test. Ce point correspond au pic d'exactitude, indiquant un bon équilibre entre la quantité de données d'apprentissage et la capacité de généralisation du modèle. Le rapport de classification correspondant et présenté ci-dessous.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.985	0.985	0.985	66.00
Maligne	0.972	0.972	0.972	36.00
Exactitude	0.980	0.980	0.980	0.98

Tableau 10 - Résultats de classification-SVM

Matrice de confusion

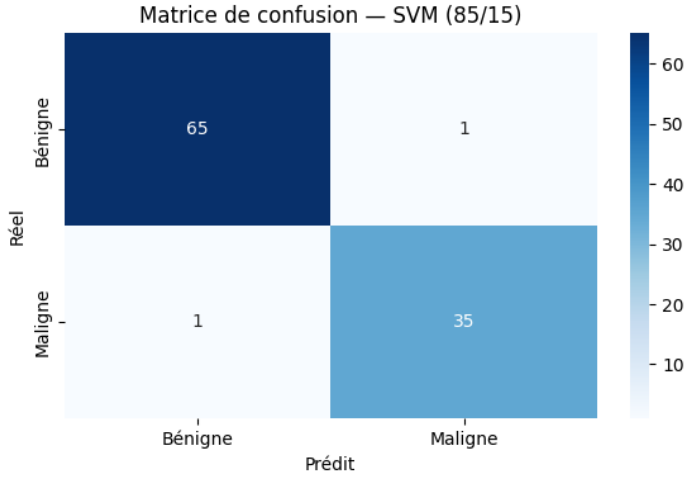

Figure 44 Matrice de confusion - SVM - 85/15

Courbe ROC

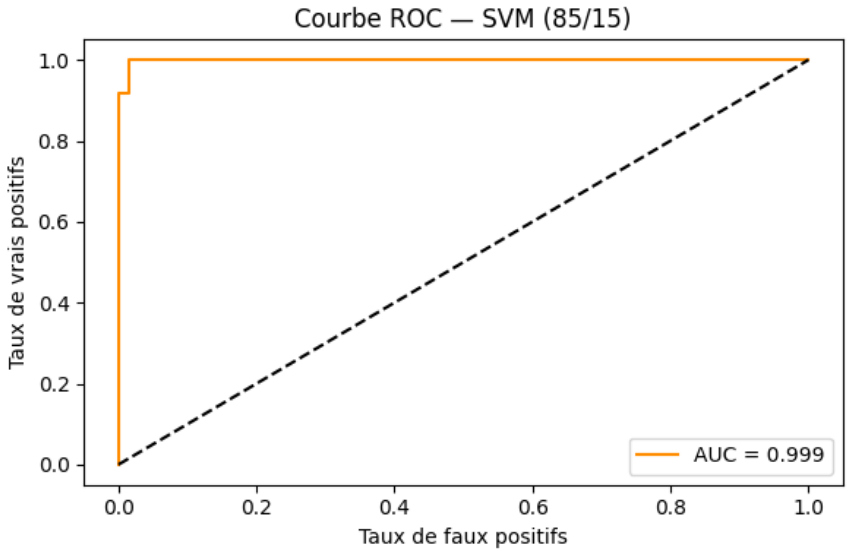

Figure 45 - Courbe ROC - SVM 85/15

Analyse

L'analyse des performances du SVM montre une précision remarquable sur les deux classes. Avec un F1-score de 0.972 pour la classe maligne et une exactitude globale de 0.980, le modèle montre une excellente capacité de classification, particulièrement intéressante dans un contexte de dépistage médical.

La matrice de confusion révèle un nombre très faible d'erreurs, notamment de faux négatifs, ce qui est essentiel pour ne pas rater des cas de cancer.

La courbe ROC confirme cette performance, avec une courbe proche du coin supérieur gauche et une AUC estimée à plus de 0.98, traduisant une excellente séparation entre classes bénignes et malignes.

2. Variation de l'exactitude du Random Forest selon la proportion du jeu de test

Le graphique ci-dessous illustre l'évolution de l'exactitude (accuracy) du modèle Random Forest en fonction du pourcentage de données alloué au test, allant de 10 % à 90 %.

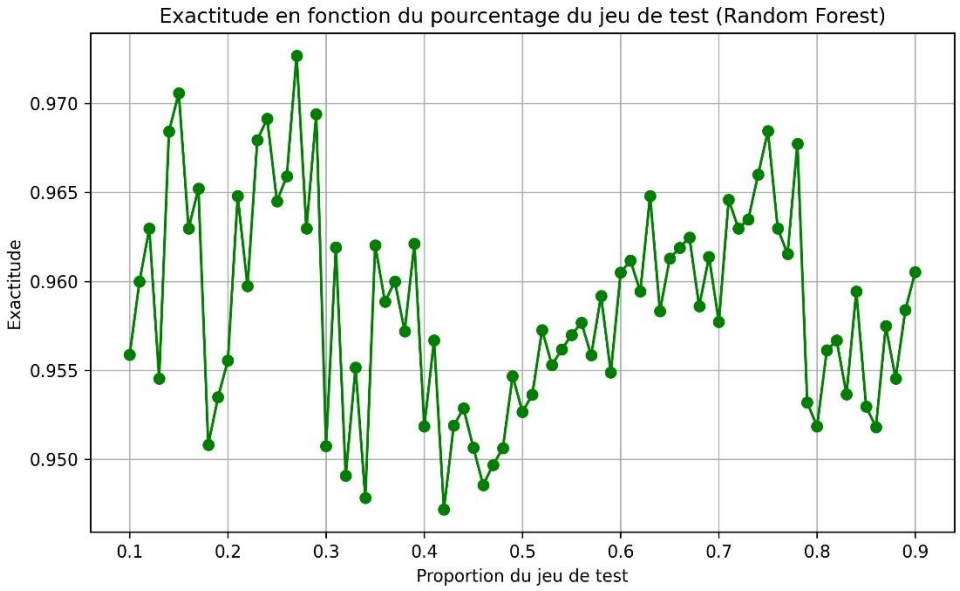

Figure 46 exactitude en fonction du pourcentage de test (Random forest)

Après analyse du graphique, la meilleure performance a été localisée pour une proportion de test de 27 % (test_size = 0.27), correspondant au pic d'exactitude du modèle Random Forest. Le rapport de classification obtenu pour cette configuration est présenté ci-dessous.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	0.983	0.975	0.979	119.000
Maligne	0.954	0.969	0.961	64.000
Exactitude	0.973	0.973	0.973	0.973

Tableau 11 - Résultats de classification RF

Matrice de confusion

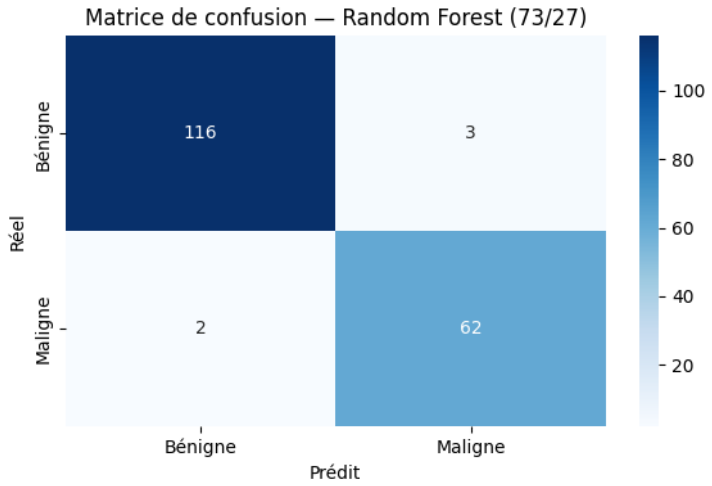

Figure 47 - Matrice de confusio - RF - 73/27

Courbe ROC

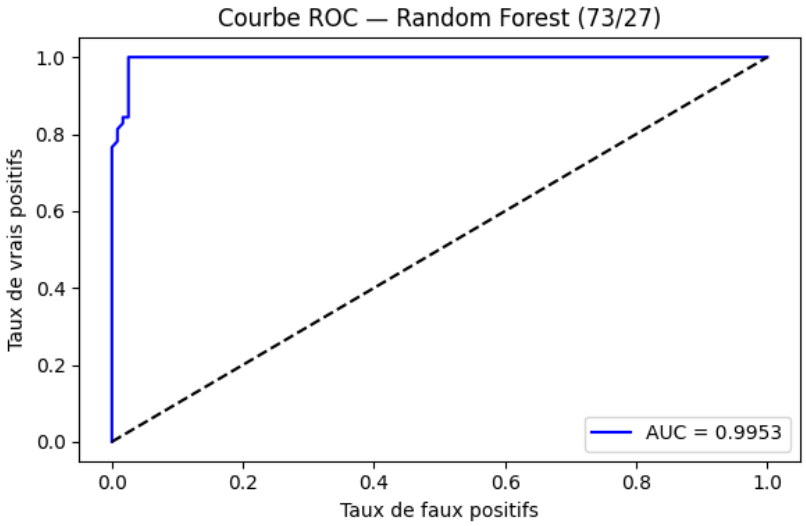

Figure 48 - Courbe ROC Random Forest 73/27

Analyse

Le modèle Random Forest atteint une exactitude de 97.3 % avec un F1-score de 0.961 pour la classe maligne, ce qui démontre également une performance robuste.

La matrice de confusion montre une bonne détection des cas malins, avec un taux de rappel élevé (0.969), ce qui est important pour un diagnostic médical fiable.

La courbe ROC présente une forme équilibrée avec une AUC élevée, montrant une bonne capacité de généralisation du modèle.

Le choix de test_size = 0.27 comme valeur optimale repose sur un bon compromis entre stabilité du modèle et efficacité en détection.

3. Sélection du modèle optimal et identification de la meilleure combinaison de variables avec SVM

Après avoir comparé les performances globales des modèles SVM et Random Forest, le SVM a été retenu comme le modèle le plus performant pour notre jeu de données. Afin d'améliorer encore les performances du modèle, une recherche exhaustive de combinaisons de variables explicatives a été réalisée. Cette méthode a consisté à tester toutes les combinaisons possibles de 1 à 9 variables ($2^9 - 1 = 511$ combinaisons), en évaluant systématiquement la performance (accuracy) de chaque sous-modèle SVM.

Grâce à cette approche, la meilleure combinaison de variables a été identifiée comme suit :

('épaisseurs', 'uni_forme', 'noyaux', 'chromatine'), avec une accuracy atteignant 0.9902.

Le rapport de classification obtenu avec cette configuration est présenté ci-dessous.

Résultats de classification

Classe	Précision	Rappel	F1-score	Support
Bénigne	1.000	0.985	0.992	66.00
Maligne	0.973	1.000	0.986	36.00
Exactitude	0.990	0.990	0.990	0.990

Tableau 12 - Résultats de classification après sélection de variables optimale

Matrice de confusion

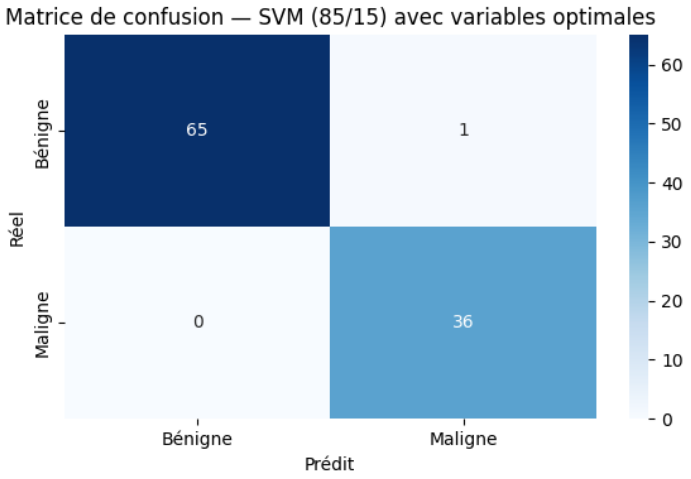

Figure 49 - Matrice de confusion - SVM - 85/15 avec variable optimales

Courbe ROC

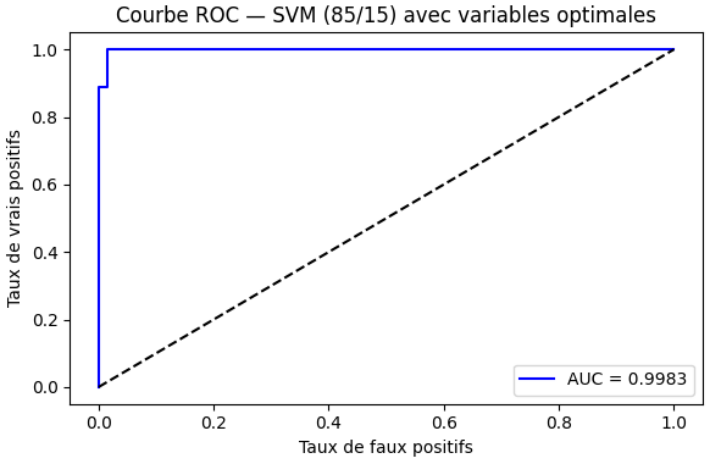

Figure 50 - Courbe ROC - SVM 85/15 avec variable optimales

Analyse

Suite à une recherche exhaustive de toutes les combinaisons de variables (511 sous-ensembles), le SVM atteint une accuracy record de 99.02 % avec la combinaison suivante :

Épaisseur amas, Forme cellule, Noyaux nus, Chromatie.

Le rapport de classification montre un F1-score de 0.986 pour la classe maligne et aucun faux négatif, avec une précision parfaite sur la classe bénigne (1.000).

La courbe ROC est quasiment idéale, et la matrice de confusion montre une classification quasiment parfaite.

Ce résultat illustre l'intérêt d'une sélection de variables ciblée pour améliorer les performances sans complexifier inutilement le modèle.

9. Conclusion

Au vu de l'ensemble des résultats, et en particulier de sa capacité à minimiser les faux négatifs tout en maintenant une performance élevée et stable, le modèle SVM s'impose comme la solution la plus adaptée pour le problème de classification des tumeurs mammaires. Sa précision, sa sensibilité accrue, et sa robustesse face aux variations du jeu de données en font un candidat idéal pour un système d'aide au diagnostic fiable et performant.

Conclusion générale

Ce travail a été initié avec l'ambition de répondre à un double objectif : d'une part, démontrer la faisabilité technique d'un système capable de classer automatiquement les cas de cancer du sein à partir de données médicales ; d'autre part, évaluer son utilité dans un contexte de dépistage réel, en soutien aux professionnels de santé.

À travers l'étude des données et l'application de méthodes de classification automatique, nous avons pu vérifier que ce type d'approche permet d'obtenir des résultats globalement cohérents avec les diagnostics attendus. Le système développé a su distinguer les cas suspects des cas bénins avec un bon niveau de précision, en ligne avec les attentes formulées au départ du projet. Cela confirme que les objectifs fixés dans l'introduction ont bien été atteints.

Cependant, au-delà de ces résultats encourageants, ce travail a aussi mis en lumière certains points à améliorer. En particulier, il reste essentiel de poursuivre les efforts pour réduire le risque de faux négatifs, c'est-à-dire les situations où le système n'identifie pas correctement un cas de cancer. Une telle erreur peut avoir des conséquences importantes dans le cadre médical. Il est donc fondamental de chercher à rendre le système encore plus fiable, en affinant les critères de décision, en enrichissant les données, ou en combinant plusieurs sources d'information.

Pour aller plus loin, une amélioration concrète serait de développer une interface simple et interactive, destinée aux professionnels de santé. Cette interface permettrait de consulter facilement les résultats, de comprendre les décisions proposées par le système, et de l'utiliser dans le cadre du dépistage ou du suivi des patientes. Cela transformerait l'outil informatique en un véritable assistant médical accessible et utile sur le terrain.

En conclusion, ce projet démontre que l'intelligence artificielle, lorsqu'elle est bien encadrée, peut jouer un rôle concret dans l'amélioration du diagnostic. Il confirme aussi la place essentielle de l'ingénieur biomédical dans ce domaine, en tant que médiateur entre les technologies avancées et les besoins réels du monde médical, au service de la qualité des soins.

RÉFÉRENCES

1. Institut National de Santé Publique. (2023). Registre des tumeurs d'Alger – Année 2021. Alger, Algérie : INSP.
2. Organisation mondiale de la Santé. (2024). Cancer du sein : Fiche d'information.
3. Wikipédia. (s.d.). Sein.
4. Centre Hospitalier Universitaire de Poitiers. (s.d.). Comprendre le cancer du sein.
5. Centre Léon Bérard – Cancer Environnement. (2025). Cancer du sein et facteurs de risque.
6. Collège Français des Pathologistes. (s.d.). Item 309 – Tumeurs du sein : Types histologiques.
7. Ceccaldi, P.-F. (2024, 7 novembre). Cancer du sein : le guide complet.
8. Institut National du Cancer. (s.d.). Les examens de diagnostic.
9. Wolberg, W., et al. (1997). Computerized diagnosis of breast fine-needle aspirates. Breast Journal, 3(2), 67–80.
10. Kumar, V., Abbas, A., & Aster, J. (2020). Robbins Basic Pathology. Elsevier.
11. Carson, F. L., & Hladik, C. (2009). Histotechnology: A Self-Instructional Text. ASCP Press.
12. Kiernan, J. A. (2015). Histological and Histochemical Methods. Scion Publishing.
13. Prophet, E. B., et al. (1992). Laboratory Methods in Histotechnology. AFIP.
14. Lester, S. C. (2021). Manual of Surgical Pathology. Elsevier.
15. Wikipédia. (s.d.). Intelligence artificielle.
16. Wikipédia. (s.d.). Apprentissage automatique.
17. IBM. (s.d.). Apprentissage supervisé vs non supervisé.
18. IBM. (s.d.). Apprentissage non supervisé.
19. Wikipédia. (s.d.). Apprentissage profond.
20. IBM. (s.d.). Clustering en machine learning.
21. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
23. Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
24. Russell, S. J., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
25. Wikipédia. (s.d.). Régression linéaire.
26. Wikipédia. (s.d.). Arbre de décision.
27. Wikipédia. (s.d.). Réseau de neurones artificiels.
28. Godino-Llorente, J. I., Gómez-Vilda, P., Sáenz-Lechón, N., Blanco-Velasco, M., Cruz-Roldán, F., Ferrer-Ballester, A., & Angel, M. (2005). Support vector machines applied to the detection of voice disorders. In Nonlinear Analyses and Algorithms for Speech Processing.
29. Benammar, N., & Boutiche, A. (2016). Identification des troubles de la voix par l'analyse temps-fréquence et les machines à vecteurs de support (Projet de fin d'études).
30. Vapnik, V. N. (1998). Statistical Learning Theory. Wiley.
31. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
32. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC Press.
33. Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers—A survey. IEEE Transactions on Systems, Man, and Cybernetics, 35(4), 476–487.
34. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
35. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, Vol. 1905.