جامعة أبو بكر بلقايد – تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

Pour l'obtention du Diplôme de

MASTER en GENIE BIOMEDICAL

Spécialité : **Informatique Biomédicale**

Présenté par : BENABDELLAH Nadir et KHALDI Diya Seyf Islem

Extraction de données exploitables pour la classification du cancer du sein et prédiction du risque de la récidive

Soutenu 07 Juillet 2021 devant le Jury

Mr	**BECHAR** Hassane	MAA	Université de Tlemcen	Président
Mme	**HEDEILI** Nawel	MAA	Université de Tlemcen	Examinatrice
Mme	**BENCHAIB** Yasmine	MCB	Université de Tlemcen	Encadreur

Année universitaire 2020-2021

Remerciement

On remercie dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mme. Benchaib Yasmine, on la remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Notre remerciement s'adresse à la doctorante Lyna Hasnaoui pour son aide pratique et son soutien moral et ses encouragements. Et au doctorant Bourega Mohamed qui nous a aidé particulièrement et suffisamment pour bien terminer notre mémoire.

Un remerciement spécial pour monsieur "Aissaoui Nabil" notre ami.

Notre remerciement s'adresse également à tous nos professeurs pour leurs générosités et la grande patience dont ils ont su faire preuve malgré leurs charges académiques et professionnelles.

J'adresse aussi mes vifs remerciements aux membres des jurys pour avoir bien voulu examiner et juger ce travail. Chaque minute pour lire notre mémoire et nous donner ses commentaires constructifs pour de meilleurs résultats, chaque remarque est important pour nous et précieuse

Que toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce mémoire trouvent ici l'expression de ma profonde gratitude.

Dédicace

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect : mon cher père Mohamed.

A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non âmes exigences et qui n'a épargné aucun effort pour me rendre heureux : mon adorable mère Alili Khayra .

Source de vie, d'amour et d'affection, que dieu les protèges.

A mes chères sœurs et leurs enfants, source de joie et de bonheur

A mes grands-mères, mes oncles et mes tantes. Source d'espoir et de motivation, Que Dieu leur donne une longue et joyeuse vie,

Spécialement à ma cousine Dr. Boucherit Ismahane

A tous mes amis, tout particulièrement

Sans oublier mon binôme Khaldi Deyaa pour son soutien moral, sa patience

et sa compréhension tout au long de ce projet

A vous cher lecteur

Dédicace

Je dédie ce mémoire :

A ma grand-mère

« Que le Tout-Puissant Allah garde leur âme dans son vaste paradis et lui donne la paix éternelle »

A mes chers parents

Pour leur encouragement et soutien moral tout au long de mon parcours universitaire

« Que Dieu leur procure la bonne santé et la longue vie »

A ma sœur, à mon petit frère

Qui m'encouragent et me soutiennent

A toute la famille, et tous mes proches

A mon binôme Benabdallah Nadir, et tous les amis.

Table des matières

Table de figures

Liste de tableaux

Résumé

La mort par cancer est l'un des problèmes majeurs de l'humanité. Bien qu'il existe de nombreux moyens de la prévenir, certains types de cancer n'ont toujours pas de traitement. L'un des types de cancer les plus courants est le cancer du sein, le diagnostic précoce et précis est l'un des choses la plus importante dans son traitement.

La récidive du cancer du sein fait partie des craintes les plus notables des femmes. La prédiction précoce de la récidive peut contribuer à apaiser ces craintes. Bien que les informations médicales soient généralement compliquées et qu'il soit difficile de simplifier les recherches pour obtenir les données les plus pertinentes, les nouvelles techniques sophistiquées d'exploration de données promettent des prédictions précises à partir de données hautement dimensionnelles.

L'objectif de ce document de recherche est de présenter un rapport sur le cancer du sein et son récidivité où nous avons profité de ces avancées technologiques disponibles pour développer des modèles de détection et de prédiction. Dans cette étude, les performances de deux algorithmes d'exploration de données établis : k-voisin le plus proche (KNN) et machine à vecteur de support (SVM), l'analyse en composantes principales (ACP), pour prédire la récurrence du cancer du sein. La comparaison a été effectuée entre les modèles construits en l'absence et en présence de l'ACP. Les résultats ont montré que KNN a produit une meilleure prédiction sans ACP (accuracy = 96.43%), tandis que l'autre technique (SVM) nous donne le même résultat avec et sans ACP. (accuracy = 78.57 %). Cette étude peut être utile au secteur de la santé en aidant les médecins à prédire précisément la récidive du cancer du sein.

Mot clés :

k- voisin le plus proche , KNN, machine à vecteur de support, SVM, l'analyse en composantes principales , ACP, Cancer de sein, récidive, récidive du cancer de sein, détection , prédiction

Abstract

Death by cancer is one of the major problems of humanity. Although there are many ways to prevent it, sometypes of cancer still have no treatment. One of the most common types of cancer is breast cancer, early and accurate diagnosis is one of the most important things in its treatment.

Recurrence of breast cancer is among the most notable fears of women. Early prediction of recurrence can help alleviate these fears. Although medical information is generally complicated and it is difficult to simplify searches to obtain the most relevant data, sophisticated new data mining techniques promise accurate predictions from highly dimensional data .

The purpose of this research paper is to report on breast cancer and its recurrence where we took advantage of these available technological advances to develop detection and prediction models. In this study, the performance of two established data mining algorithms: k-nearest neighbor (KNN) and support vector machine (SVM), principal component analysis (PCA), to predict breast cancer recurrence. Comparison was made between models built in the absence and presence of PCA. The results showed that KNN produced a better prediction without PCA (accuracy = 96.43%), while the other technique (SVM) gave us the same result with and without PCA. (accuracy = 78.57%). This study can be useful to the health sector by helping physicians toaccurately predict breast cancer recurrence.

Keyword:

k-nearest neighbor , KNN, support vector machine, SVM, principal component analysis , PCA, breast cancer, recurrence, breast cancer recurrence, detection , prediction.

الملخص

الموت من السرطان هو أحد المشاكل الرئيسية للبشرية. على الرغم من وجود العديد من الطرق للوقاية منه، إلا أن بعض أنواع السرطان لا تزال بلا علاج. يعد سرطان الثدي من أكثر أنواع السرطانات شيوعًا ، ويعد التشخيص المبكر والدقيق من أهم الأمور في علاجه.
من أبرز مخاوف النساء تكرار الإصابة بسرطان الثدي, يمكن للتنبؤ المبكر أن يساعد في تهدئة هذه المخاوف. على الرغم من أن المعلومات الطبية معقدة بشكل عام ومن الصعب تبسيط عمليات البحث للحصول على البيانات الأكثر صلة ، فإن تقنيات التنقيب عن البيانات الجديدة والمعقدة تعد بتنبؤات دقيقة.
الهدف من هذا البحث هو تقديم تقرير عن سرطان الثدي وتكرار حدوثه حيث استفدنا من هذه التطورات التكنولوجية المتاحة لتطوير نماذج للكشف والتنبؤ . في هذه الدراسة ، تم أداء خوارزميتين لاستخراج البيانات) KNN (:،) SVM (تحليل المكون الرئيسي) PCA (و للتنبؤ بتكرار الإصابة بسرطان الثدي.
تم إجراء المقارنة بين النماذج التي تم إنشاؤها في غياب وحضور PCA ،أظهرت النتائج أن KNN أنتجت تنبؤًا أفضل بدون96.43) PCA ٪)، بينما تعطينا التقنية الأخرى (SVM) نفس النتيجة مع PCA وبدونه) 78.57 ٪(. قد تكون هذه الدراسة مفيدة للرعاية الصحية من خلال مساعدة الأطباء على التنبؤ بدقة بتكرار الإصابة بسرطان الثدي.

كلمات مفتاحية :
تحليل المكون الرئيسي ، SVM، آلة ناقلات الدعم ، KNN,الجار الأقرب -k، سرطان الثدي ، تكرار ، تكرار سرطان الثدي ، الكشف ، PCA تنبؤ.

Introduction générale

La mort par cancer est l'un des problèmes majeurs de l'environnement sanitaire. Le cancer du sein est le type de cancer le plus fréquent chez les femmes dont le tissu mammaire est plus dense en raison de ses caractéristiques physiologiques et l'une des raisons les plus fréquents de la mort chez les femmes.[1]

En 2020, il y avait 2.3 millions de femmes diagnostiquées avec un cancer du sein et 685000 décès dans le monde [2]. À la fin de l'année 2020, 7.8 millions de femmes vivantes ont reçu un diagnostic de cancer du sein au cours des cinq dernières années, ce qui fait ce cancer le plus répandu au monde. Le nombre d'années de vie ajusté à l'incapacité (DALYs) perdues par les femmes à cause du cancer du sein dans le monde est supérieur à celui de tout autre type de cancer [2]. Il est présent dans tous les pays du monde, chez les femmes à tout âge après la puberté, mais avec des taux croissants à un âge plus avancé.[2]

Donc, la détection de cette maladie à un stade précoce permet d'éviter l'augmentation du nombre de décès. Il est important de poser un diagnostic précis des tumeurs. La plupart des tumeurs sont le résultat de changements bénins (non cancéreux) dans le sein, mais si une tumeur maligne est diagnostiquée comme bénigne, elle causera de graves problèmes.

La détection précoce du cancer du sein et l'obtention d'un traitement moderne sont les stratégies les plus importantes pour prévenir les décès dus au cancer du sein. Il est facile de traiter avec succès un cancer du sein précoce, de petite taille et qui ne se propage pas. Le moyen le plus sûr de détecter un cancer du sein à un stade précoce est de passer régulièrement des tests de dépistage. Malgré l'augmentation du nombre d'études médicales et les développements technologiques qui contribuent au traitement du cancer, il existe encore des problèmes dans le diagnostic du cancer.

Après le diagnostic et le traitement, La récidive du cancer du sein est l'un des plus grands défis auxquels une patiente doit faire face et l'un des problèmes qui ont un impact sur son niveau de vie. La récidive du cancer du sein désigne la réapparition d'un cancer du sein chez une femme dont l'ancien cancer a été guéri. Des études montrent que même 20 ans après le diagnostic, les femmes atteintes d'un type de cancer du sein alimenté par les œstrogènes sont toujours confrontées à un risque important de récidive ou de propagation du cancer.[3]

La prédiction est un défi car les données sur la récidive sont rarement enregistrées dans la plupart des ensembles de données sur le cancer du sein. Une prédiction précise et opportune est essentielle car elle aide les médecins à prendre une décision et favorise une thérapie plus personnalisée des patients.

Dans ce document, nous discutons la classification et la prédiction de cancer du sein à travers trois principaux chapitres.

Le premier chapitre contient la définition du cancer de sein, certaines connaissances (le dépistage et diagnostique, traitement) et le risque de la récidivité. Après en passant au deuxième chapitre qui concerne les méthodes d'extraction de connaissance à partir des données (les définitions des méthodes utilisées dans la classification et la prédiction) et la méthode Analyse en Composant Principale (ACP) de sélection des paramètres. Dans le troisième chapitre, Nous proposons notre solution, les étapes suivies (définition et la préparation des bases et l'utilisation de KNN et SVM) et à la fin nous avons interprété les résultats obtenus.

Chapitre I

Cancer du sein

1. Introduction

Le cancer du sein représente un problème majeur de santé publique, d'un grand intérêt et d'une grande importance chez les médecins de diverses spécialités et chez les femmes en général[1]. Il est la cause la plus fréquente du décès chez les femmes dans le monde. Le cancer du sein a dépassé le cancer du poumon en tant que cancer le plus souvent diagnostiqué, les taux varient environ cinq fois dans le monde, mais ils augmentent dans les régions qui, jusqu'à récemment, présentaient de faibles taux de maladie[2]. Cependant, en Algérie reste un des rares pays où le cancer chez la femme dépasse celui de l'homme. Dans Le Monde, Le cancer du sein est une principale cause de décès avec environ 69.000 par ans. Dans L'Algérie, le Pr. Kamel Bouzid, indiquant que 14.000 nouveaux cas du cancer du sein sont enregistrés chaque année en Algérie.

Pour cela et en raison du rôle central du sein dans la production du lait pour les nouveau-nés et l'image de la femme, il est important de bien comprendre le cancer du sein et son comportement. Ce chapitre s'organisera en trois points principaux :

- L'anatomie et la physiologie du sein.
- Définition des notions élémentaires sur le cancer du sein.
- Une bonne partie du chapitre sera consacrée pour description des phases du dépistage, diagnostique et traitement de cette maladie.

2. Le sein

Les seins jouent un rôle important dans la féminité et dans l'image que la femme a de son corps. La fonction biologique du sein est de produire du lait afin de nourrir un nouveau-né.

Chaque sein contient une glande mammaire (elle -même composée de quinze à vingt compartiments séparés par du tissu graisseux) qui donne au sein la forme qu'on lui connaît et du tissu de soutien qui contient des vaisseaux, des fibres et de la graisse.

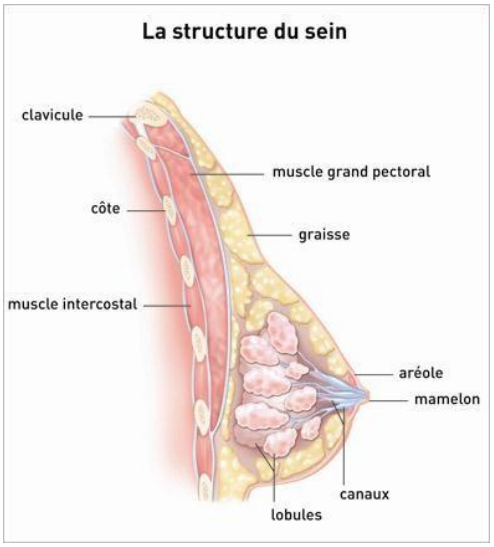


Figure 1: Anatomie du sein.

Chacun des compartiments de la glande mammaire est constitué de lobules et de canaux. Le rôle des lobules est de produire le lait en période d'allaitement. Les canaux transportent le lait vers le mamelon.

La glande mammaire se développe et fonctionne sous l'influence des hormones sexuelles fabriquées par les ovaires. Ces hormones sont de deux types :

- Les œstrogènes, qui permettent notamment le développement des seins au moment de la puberté et jouent un rôle important tout au long de la grossesse (assouplissement des tissus, augmentation du volume sanguin nécessaire à l'alimentation du bébé, etc.).
- La progestérone qui joue notamment un rôle dans la différentiation des cellules du sein et sur le cycle menstruel, en préparant par exemple l'utérus à une éventuelle grossesse (densification et développement de la vascularisation la muqueuse de l'utérus).

Le sein est parcouru de vaisseaux sanguins et de vaisseaux lymphatiques. Les ganglions et les vaisseaux lymphatiques composent le système lymphatique qui aide notamment à combattre les infections. Les ganglions lymphatiques du sein sont principalement situés :

- Au niveau de l'aisselle (ganglions axillaires) ;
- Au-dessus de la clavicule (ganglions sus-claviculaires); sous la clavicule (ganglions sous-claviculaires ou infra-claviculaires) ;
- A l'intérieur du thorax, autour du sternum (ganglions mammaires internes).[3]

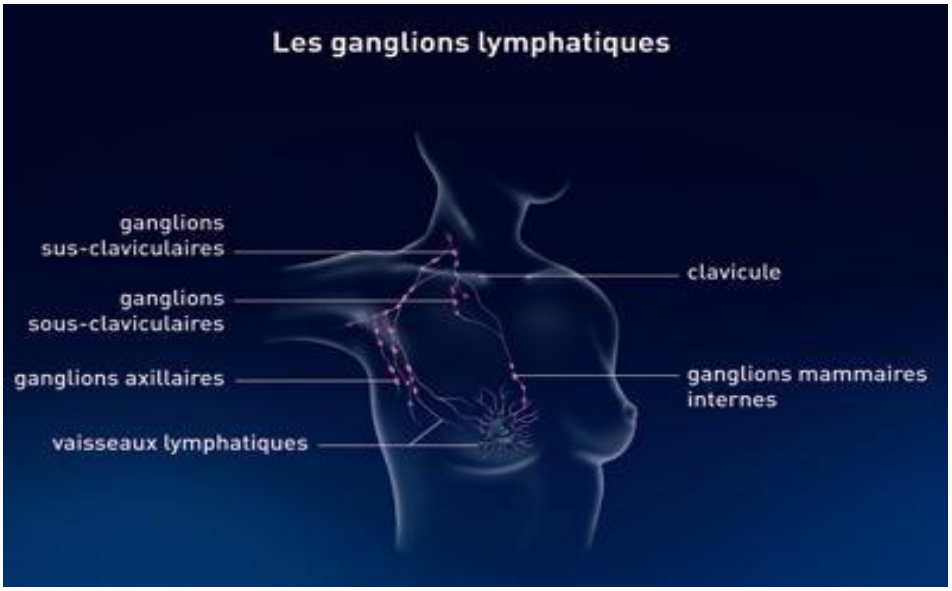


Figure 2 :les ganglions lynphatiques.

2.1. Cancer du sein

2.2. Définition

Le cancer du sein est une maladie connue depuis l'antiquité dans le bassin méditerranéen et ses premières descriptions remontent à l'Égypte ancienne, environ 2500 ans avant JC.[4]

Un cancer du sein est une tumeur maligne qui touche la glande mammaire, il résulte d'un dérèglement de certaines cellules qui se multiplient et forment le plus souvent une masse appelée tumeur.

On parle alors d'adénocarcinome. "Une cellule de la glande mammaires se transforme et se développe soudainement de manière anarchique ; les cellules qui en résultent, prolifèrent sans s'arrêter et peuvent migrer dans d'autres parties du corps. Ce sont alors des métastases, qui peuvent atteindre les os, les organes comme le foie ou les poumons…", explique le Pr Jean-Yves Pierga. Ces cellules se développent à partir de l'épithélium qui borde les canaux galactophores (dans lesquels circule le lait).[5]

Les cellules peuvent rester dans le sein ou se répandre dans le corps par les vaisseaux sanguins ou lymphatiques. La plupart du temps, la progression d'un cancer du sein prend plusieurs mois et même quelques années.

Dans 95 % des cas, les cancers du sein sont des adénocarcinomes. Ils se développent le plus souvent à partir des cellules des canaux (cancer canalaire) et plus rarement à partir des cellules des lobules (cancer lobulaire).[6]

2.3. Symptômes

Il est possible que le cancer du sein ne cause aucun signe ni symptôme aux tout premiers stades de la maladie. Les symptômes apparaissent quand la tumeur au sein est suffisamment grosse pour qu'on sente la masse au toucher ou quand le cancer s'est propagé aux tissus et organes voisins. D'autres affections médicales peuvent causer les mêmes symptômes que le cancer du sein.

Le symptôme le plus fréquent du carcinome canalaire est une masse ferme ou dure qui est très différente du reste du tissu mammaire. Elle peut sembler fixée à la peau ou au tissu mammaire voisin. La masse ne rétrécit pas ou ne disparaît pas et ne réapparaît pas au cours du cycle menstruel. Elle peut être sensible mais n'est généralement pas douloureuse. (La douleur est plus souvent le symptôme d'une affection non cancéreuse.)

Il arrive souvent que le carcinome lobulaire ne forme pas de masse. On a plus l'impression que le tissu mammaire s'épaissit ou durcit.

Les autres symptômes du cancer du sein canalaire ou lobulaire peuvent être ceux-ci :

- masse à l'aisselle (creux axillaire)
- changement de la taille ou de la forme du sein
- changements mamelonnaires, comme un mamelon qui commence soudainement à pointer vers l'intérieur (mamelon inversé)
- écoulement du mamelon sans qu'on le comprime ou qui est teinté de sang.

Les signes et symptômes tardifs se manifestent quand la masse cancéreuse grossit ou se propage à d'autres parties du corps, dont d'autres organes :

- douleur osseuse
- perte de poids
- nausées

- perte d'appétit jaunisse.
- essoufflement
- toux
- maux de tête
- vision double
- faiblesse musculaire

Le cancer inflammatoire du sein et la maladie de Paget du sein causent des symptômes différents.[12]

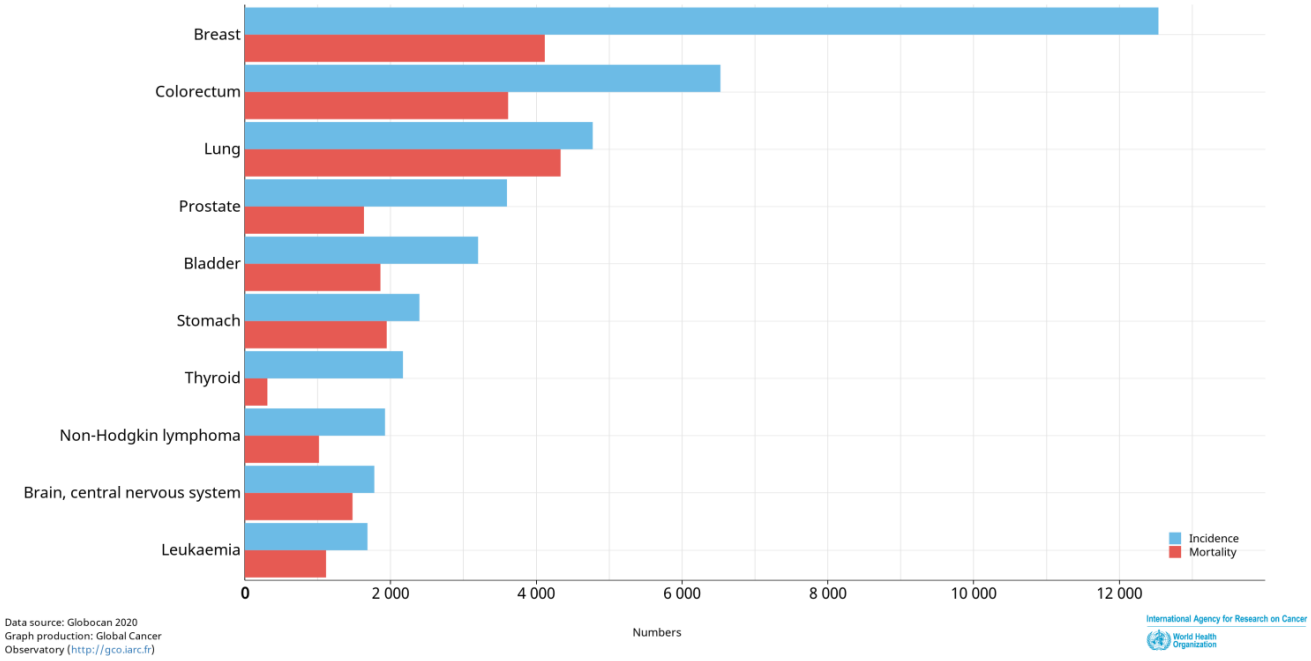


Figure 3: Le nombre estimé des cas de mort en Algérie, tous les sexes, tous les ages.

2.3.1. **Dépistage de cancer du sein :**

a. **Le dépistage individuel :**

Le cancer du sein est le cancer le plus fréquent chez la femme dans le monde. Il constitue également la principale cause de mortalité par cancer chez les femmes. Plusieurs actions peuvent être mises en place afin de favoriser une détection précoce du cancer du sein. L'intérêt est de pouvoir soigner ce cancer plus facilement et de limiter les séquelles liées à certains traitements.

Plusieurs actions peuvent être mises en place afin de favoriser une détection précoce de ce cancer :

1. La consultation d'un médecin en cas de changements au niveau des seins :
    - Apparition d'une boule, d'une grosseur dans le sein ou sous un bras (aisselle).
    - Modification de la peau (rétraction, rougeur, œdème ou aspect de peau d'orange).
    - Modification du mamelon ou de l'aréole - zone qui entoure le mamelon - (rétraction, changement de coloration, suintement ou écoulement).
    - Changements de forme des seins.

Ces signes ne signifient pas nécessairement la présence d'un cancer mais doivent être signalés au médecin[11].

2. un examen clinique des seins (palpation) :

Les recommandations indiquent que l'on doit orienter les patientes à risque dès l'âge de 25 ans vers une personnalisation du dépistage. Les consultations gynécologiques sont l'occasion de se dépister, grâce à la palpation mammaire[12].Cet examen rapide et indolore permet de détecter une éventuelle anomalie. Il peut être réalisé par un généraliste, un gynécologue ou une sage-femme [11] Les femmes peuvent également pratiquer l'autopalpation afin de détecter une éventuelle masse.

3. Comment s'auto-palper les seins ?

L'autopalpation est un geste qu'il est conseillé d'effectuer tous les mois après les règles, afin de repérer une éventuelle grosseur du sein. Mettez-vous d'abord debout devant un miroir, inspectez les deux seins et vérifiez l'absence d'écoulement d'un mamelon, de crevasses, de plis anormaux ou d'une peau qui pèle. Puis, levez un des deux bras, puis avec les 3 doigts de l'autre main, palpez le sein du côté du bras levé : débutez par la partie externe, les doigts à plat, en effectuant de petits cercles. Il faut rechercher toute sensation de boule, de fossette sur la peau, ou de grosseur. Palpez également le mamelon et la zone entre le sein et l'aisselle, pressez le mamelon et vérifiez qu'aucun écoulement ne se produise[12].

b. Le dépistage organisé

Dès 50 ans, le dépistage individuel laisse la place au dépistage organisé, une mammographie de dépistage (examen radiologique) associée à un examen clinique des seins, proposée tous les deux ans aux femmes de 50 à 74 ans en l'absence de symptôme apparent ou de facteur de risque. Parce que près de 80% des cancers du sein touchent des personnes âgées de plus de 50 ans. Des examens complémentaires (échographie, IRM, biopsie...) peuvent être proposés si nécessaire. Dans le cadre du programme de dépistage organisé, mis en place depuis 2004 pour les femmes de 50 à 74 ans, les mammographies jugées normales font l'objet d'une seconde lecture systématique, par sécurité, assurée par un autre radiologue expert. Comme tout acte médical, le dépistage du cancer du sein présente des bénéfices et des limites, qu'il est important de connaître avant de prendre une décision [11].

Si vous présentez un facteur de risque particulier (antécédents de cancer ou plusieurs cas de cancers du sein dans votre famille), et ce quel que soit votre âge, adressez-vous à votre médecin traitant ou à votre gynécologue qui vous proposera une modalité de surveillance plus spécifique.

Détecté à un stade précoce, le cancer du sein peut être guéri dans plus de 90 % des cas. Le moyen de détection est la mammographie qui permet de dépister, avant tout symptôme, 90 % des cancers du sein déjà présents[13].

c. La mammographie

Une mammographie est une radiographie des seins. Dans le cadre du dépistage, elle permet notamment de détecter des cancers de petite taille, bien avant qu'ils ne soient palpables ou que des symptômes n'apparaissent.

Elle est réalisée avec un appareil de radiologie spécifique appelé mammographie qui utilise les Rayons X. L'un après l'autre, vos seins sont placés entre deux plaques qui se resserrent et compriment le sein pendant quelques secondes. Deux clichés par sein sont réalisés.

Le médecin interprète immédiatement les clichés et effectue ensuite un examen clinique ; il s'agit d'une palpation de votre poitrine pour repérer certaines anomalies parfois non détectables à la mammographie. Un entretien avec le médecin complète cet examen [13].

2.3.2. **Diagnostique**

L'examen clinique comprend l'inspection puis la palpation des seins et des aires ganglionnaires. Vient par la suite de la mammographie en cas de signes d'appel cliniques ou dans le cadre d'un dépistage. Elle peut être complétée par un certain nombre d'examens afin de caractériser plus précisément la nature d'une lésion repérée par la mammographie notamment dans les seins denses.

b. Echographie Mammaire

Dans le cadre d'un diagnostic de cancer du sein, l'échographie est particulièrement adaptée pour les lésions qui sont déjà, repérées par une mammographie (les kystes), pour mesurer leurs tailles et voir la nature, liquide ou solide, des nodules palpés ou découverts, elle est donc en deuxième intention après la mammographie. Une échographie complète certaines mammographies difficiles à interpréter, en cas de seins denses rendant son analyse très difficile car pouvant masquer de petites lésions ayant la même densité que le tissu mammaire environnant. Elle utilise des ultrasons pour produire des images de l'intérieur du sein. Cet examen est pratiqué par un radiologue.

c. Imagerie par Résonance Magnétique IRM

L'imagerie par résonance magnétique du sein reste un examen de troisième intention en aval de la mammographie et de l'échographie, réservé à des populations à risque accru de cancer mammaire (cas d'antécédent familial de cancer du sein ou d'antécédent d'irradiation thoracique médicale à haute dose), ou dans le cas d'une rupture intra-capsulaire d'une prothèse mammaire. Elle proposant une visualisation de coupes anatomiques extrêmement fines, elle permet de différencier les tissus pathologiques des tissus sains et de réaliser une "cartographie" très précise et en trois dimensions des tumeurs. Il reste optionnel en présence d'une forte densité mammaire, elle n'est pas préconisée en première intention en cas de masse palpable.

d. L’aspiration ou ponction cytologique

Une ponction cytologique consiste à prélever des cellules au niveau d’une anomalie du sein. Les cellules sont analysées au microscope afin d’identifier la nature de la lésion, de décider si un traitement est nécessaire et si c’est le cas, d’orienter les médecins sur le choix du traitement.

Les cellules sont prélevées à travers la peau (prélèvement percutané) à l’aide d’une seringue et d’une aiguille fine. On parle aussi de ponction à l’aiguille fine. Plus rarement, lorsque la lésion dans le sein évoque un cancer, cette ponction est également effectuée dans des ganglions de l’aisselle.

Peu douloureuse et rapide, cette technique ne nécessite pas d’anesthésie locale ni d’hospitalisation ; elle est effectuée par le gynécologue, le radiologue ou le chirurgien[14].

2.3.3. La biopsie

La biopsie est le seul examen qui permet de confirmer un diagnostic de cancer du sein ; elle est réalisée sous anesthésie locale. Lors de l’examen, le médecin utilise une aiguille fine avec laquelle il pique la peau au niveau du sein atteint. En se guidant grâce à une sonde d’échographie ou sous scanner, il prélève un échantillon du tissu anormal. Cet échantillon est ensuite analysé sous microscope afin de confirmer ou non la nature cancéreuse de la lésion et son degré d’extension local (in situ ou infiltrant). Contrairement aux ponction cytologiques qui permettent de prélever des liquides, les biopsies enlèvent des fragments du tissu mammaire[15].

2.3.4. Traitements

a. La chirurgie

Le traitement des cancers par chirurgie consiste à **retirer la tumeur**. On parle d’exérèse ou de résection. Elle est utilisée dans environ 80 % des cas : **sa visée est curative** (lorsqu’elle permet de retirer 100 % des cellules tumorales). Mais elle peut également être réalisée à **visée diagnostique** (le tissu retiré est analysé pour préciser la nature de la lésion et faciliter le choix de traitements complémentaires à visée curative) ou **palliative** (pour soulager la douleur liée à la taille de la tumeur, pour faciliter le fonctionnement de l’organe atteint...)[16].

Deux types d’interventions chirurgicales peuvent être pratiqués : une chirurgie mammaire conservatrice, appelée tumorectomie ou segmentectomie ou une chirurgie mammaire non conservatrice, appelée mastectomie.

- **La chirurgie conservatrice (ou tumorectomie ou segmentectomie) :** consiste à retirer la tumeur et une petite quantité des tissus qui l’entourent de façon à conserver la plus grande partie de votre sein. Elle est privilégiée aussi souvent que possible, en concertation avec vous. Elle est toujours complétée d’une radiothérapie .l’usage de cette technique tend à augmenter actuellement, cependant, elle comporte un risque accru de récidive locale, particulièrement chez la femme jeune ; c’est pour cela qu’elle doit être suivie de radiothérapie afin de limiter ces récidives.
- **La chirurgie non conservatrice (ou mastectomie) :** consiste à retirer la totalité du sein y compris l'aréole et le mamelon. Dans ce cas, différentes techniques de reconstruction du sein peuvent vous être proposées [17].

b. La radiothérapie

La radiothérapie se fonde sur l'utilisation de rayons ionisants dont la forte énergie permet de détruire les cellules cancéreuses. Deux types de radiothérapie existent : la radiothérapie externe et la radiothérapie interne (ou curiethérapie).

- **La radiothérapie externe :** les rayons thérapeutiques sont émis par une source externe placée au regard de la lésion. Ils traversent la peau du patient pour atteindre leur objectif.
- **La radiothérapie interne :** les rayonnements sont émis par une source qui est introduite sur le site même de la tumeur. Il s'agit en règle générale de billes, de microsphères ou de fils composés d'iridium ou de césium radioactif.

La radiothérapie, seule ou en association avec la chimiothérapie, est généralement à visée curative. Elle est parfois utilisée comme un traitement palliatif, pour diminuer les symptômes locaux associés à la tumeur.[16]

c. La chimiothérapie

La chimiothérapie passe par l'administration de médicaments dits « cytotoxiques » qui vont détruire les cellules tumorales. Ces médicaments peuvent agir sur différents processus impliqués dans la multiplication des cellules. Un protocole de chimiothérapie fait souvent appel à une association de plusieurs médicaments qui agissent sur ces différents processus. Ils sont administrés quotidiennement ou par cures, avec une fréquence variable. Chaque cure consiste à traiter le patient pendant plusieurs jours, puis à observer une période de repos durant laquelle les cellules saines peuvent se régénérer.

Les chimiothérapies sont souvent redoutées en raison de leurs effets secondaires (chute des cheveux, nausées, vomissements, baisse du nombre de cellules sanguines...). En effet, les médicaments de chimiothérapie s'attaquent non seulement aux cellules tumorales mais aussi aux cellules saines qui se multiplient activement comme celles des cheveux, du sang ou des muqueuses digestives[16].

d. L'hormonothérapie

La croissance de certains cancers est favorisée par les hormones sexuelles produites par l'organisme : ainsi, certaines tumeurs du sein ou de l'utérus croissent sous l'action des œstrogènes ou de la progestérone, et certains cancers de la prostate progressent sous l'action de la testostérone. Les hormones sexuelles agissent sur les cellules tumorales en se fixant à leur surface au niveau de récepteurs spécifiques. Les médicaments d'hormonothérapie bloquent la synthèse de ces hormones ou empêchent leur fixation aux récepteurs.

Avant de démarrer un traitement par hormonothérapie, les médecins doivent s'assurer que le cancer est bien « hormono-dépendant », c'est-à-dire que sa croissance dépend de l'action d'hormones sexuelles. Pour cela, ils recherchent la présence de récepteurs aux hormones sur les cellules tumorales. Cette vérification passe par une analyse moléculaire, conduite à partir d'échantillons de la tumeur[16].

e. Les thérapies ciblées

Les thérapies ciblées constituent une autre famille de traitements innovants du cancer. Il s'agit de médicaments qui s'attaquent spécifiquement aux cellules cancéreuses en reconnaissant des structures ou des fonctions qui leur sont propres.

D'autres médicaments de thérapies ciblées agissent sur l'environnement des cellules cancéreuses, par exemple en bloquant la formation des vaisseaux sanguins qui irriguent la tumeur. On parle alors de médicaments anti-angiogéniques.

L'introduction des thérapies ciblées a constitué une véritable révolution dans la prise en charge de certains cancers : ainsi, grâce à un médicament nommé "imatinib", les leucémies myéloïdes chroniques (LMC) qui étaient auparavant mortelles sont aujourd'hui devenues pour de nombreux patients des maladies chroniques. Dans les cancers du sein sur exprimant le récepteur HER2, le pronostic de la maladie a été significativement amélioré par la découverte du trastuzumab, une molécule de thérapie ciblée qui bloque le fonctionnement de ce récepteur [16].

2.3.5. Risque de Récidive :

Le cancer du sein récurrent est un cancer du sein qui réapparaît après le traitement initial. Bien que le traitement initial vise à éliminer toutes les cellules cancéreuses, quelques-unes ont peut-être échappé au traitement et ont survécu. Ces cellules cancéreuses non détectées se multiplient, devenant un cancer du sein récurrent.

Le cancer du sein récurrent peut survenir des mois ou des années après votre traitement initial. Le cancer peut revenir au même endroit que le cancer d'origine (récurrence locale), ou il peut se propager à d'autres parties de votre corps (récurrence à distance).

Apprendre que vous avez un cancer du sein récurrent peut être plus difficile que d'agir avec le diagnostic initial. Mais avoir un cancer du sein récurrent est loin d'être sans espoir. Le traitement peut éliminer le cancer du sein récurrent local, régional ou éloigné. Même si un traitement n'est pas possible, le traitement peut contrôler la maladie pendant de longues périodes[18].

2.3.6. Symptômes

Les signes et les symptômes[18] du cancer du sein récurrent varient selon l'endroit où le cancer revient.

a. Récurrence locale

Lors d'une récidive locale, le cancer réapparaît dans la même région que votre cancer initial.

Si vous avez subi une tumorectomie, le cancer pourrait réapparaître dans le tissu mammaire restant. Si vous avez subi une mastectomie, le cancer pourrait réapparaître dans le tissu qui tapisse la paroi thoracique ou dans la peau.

Les signes et symptômes d'une récidive locale au sein d'un même sein peuvent comprendre :

- Une nouvelle grosseur dans la poitrine ou une zone de fermeté irrégulière
- Changements dans la peau de votre sein
- Inflammation de la peau ou zone de rougeur
- Écoulement du mamelon

Les signes et symptômes d'une récidive locale sur la paroi thoracique après une mastectomie peuvent inclure :

- Un ou plusieurs nodules indolores sur ou sous la peau de la paroi thoracique
- Une nouvelle zone d'épaississement le long ou à proximité de la cicatrice de mastectomie

b. Récurrence régionale

Une récidive régionale du cancer du sein signifie que le cancer est revenu dans les ganglions lymphatiques voisins.

Les signes et symptômes d'une récidive régionale peuvent inclure une bosse ou un gonflement des ganglions lymphatiques situés :

- Sous votre bras
- Près de la clavicule
- Dans la rainure au-dessus de la clavicule
- Dans ton cou

c. Récurrence à distance

Une récurrence lointaine (métastatique) signifie que le cancer a voyagé vers des parties éloignées du corps, le plus souvent les os, le foie et les poumons.

Les signes et symptômes comprennent :

- Douleurs persistantes et qui s'aggravent, comme des douleurs à la poitrine, au dos ou à la hanche
- Toux persistante
- Difficulté à respirer
- Perte d'appétit
- Perte de poids sans efforts
- Maux de tête graves
- Crises

Quand consulter un médecin ?

Après la fin de votre traitement contre le cancer du sein, votre médecin établira probablement un calendrier d'examens de suivi pour vous. Pendant les examens de suivi, votre médecin vérifie s'il y a des symptômes ou des signes de récidive du cancer.

Vous pouvez également signaler tout nouveau signe ou symptôme à votre médecin. Prenez rendez-vous avec votre médecin si vous remarquez des signes et des symptômes persistants qui vous inquiètent[18].

3. **Conclusion**

Dans ce chapitre, nous avons introduit quelques notions générales concernant l'exploration du cancer du sein, à travers un bilan sénologique (anatomie et physiologie du sein). L'accent a été mis sur le diagnostic de ce cancer ainsi que les possibilités du traitement.

Due à la complexité de cette pathologie qui représente selon l'OMS, la cause principale du décès chez les femmes et dans le but d'augmenter les chances de guérison totale, nous représenterons dans les chapitres suivants, le processus et les méthodes utilisées pour l'implémentation d'un modèle automatique intelligent d'aide au diagnostic, détection et prédiction de la récidivité du cancer du sein.

Chapitre II

Méthodes d'ECD et de classification

1. Introduction

L'extraction de connaissances à partir de bases de données (E.C.B.D.) est une discipline émergente à l'intersection des bases de données, de l'intelligence artificielle, des statistiques, des interfaces homme-machine et de la visualisation. A partir des données collectées par les experts, le but est de fournir de nouvelles connaissances qui enrichissent l'interprétation du champ d'application, et en même temps de fournir une méthode automatique pour utiliser ces informations. ..) . Ce chapitre s'organisera en deux (02) points principaux :

- Les processus de l'extraction de données.
- Les généralités de fouilles de données.

2. L'Extraction des Connaissances à partir des Données (ECD)

2.1. Processus d'ECD

Le processus d'extraction de connaissances dans les bases de données (ECD) est présenté dans la figure suivante :

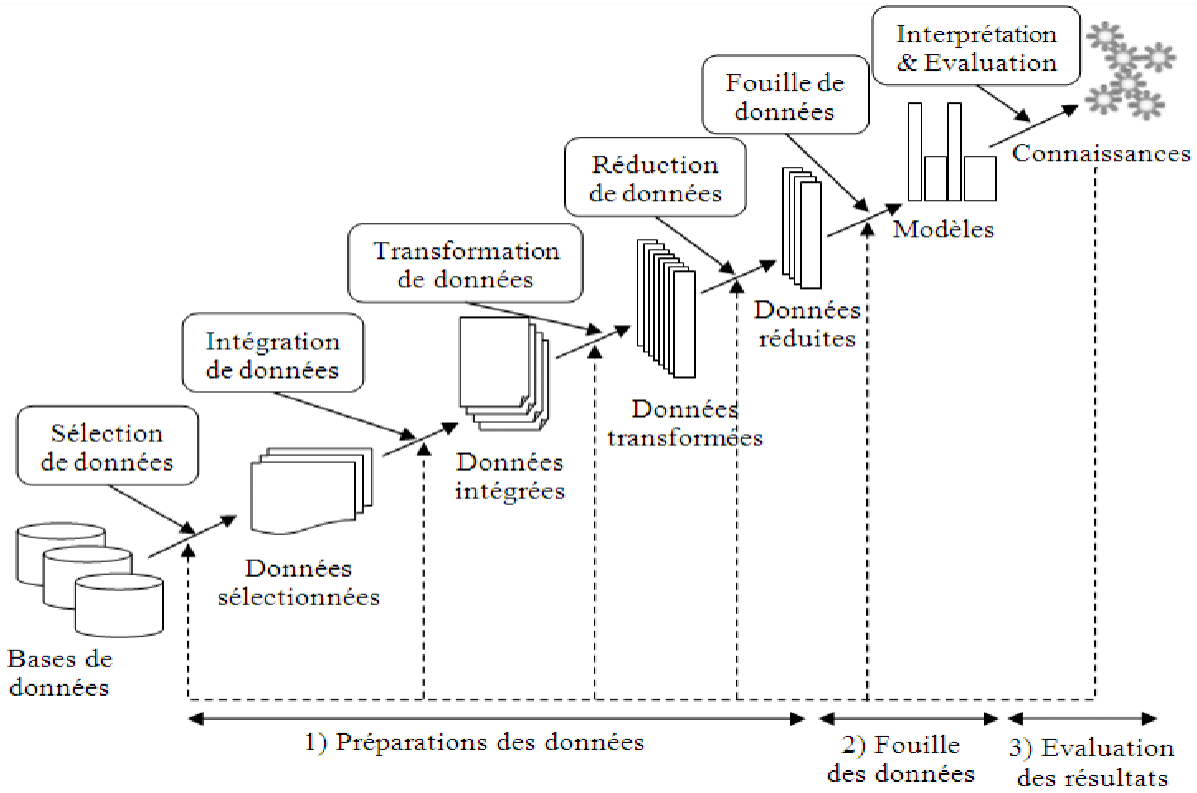


Figure 4: Les différentes étapes du processus d'ECD.

Le processus ECD peut avoir deux objectifs, soit valider l'hypothèse de l'utilisateur, soit tester l'hypothèse de l'utilisateur. [19]

2.2. Etapes d'un processus d'extraction de connaissance :

Il s'agit de quatre (04) étapes principales[19]

a. Traitement de données

Le traitement de données consiste à traiter les données bruitées pour faciliter l'exploitation dans le futur.

b. Préparation des données

Cette étape permet de sélectionner et de transformer des données afin qu'elles puissent être exploitées par des outils de data mining (fouille de données).

c. Fouille de données

La fouille de données (*data mining* en anglais), est le moteur du processus d'ECD. Le but de cette méthode c'est extraire les connaissances d'apprît des méthodes intelligentes.

d. Interprétation

Cette phase consiste l'évaluation et la présentation des résultats.

2.2.1. Fouille de données (Data mining)

Connue aussi sous l'expression de l'exploitation de donnée ou extraction de connaissances à partir de données, Il vise à extraire des connaissances à partir de grandes quantités de données par des méthodes automatiques ou semi-automatiques. [19]

Il propose un ensemble d'algorithmes dans diverses disciplines scientifiques telles que les statistiques, l'intelligence artificielle ou l'informatique pour construire des modèles à partir de données, c'est-à-dire trouver des structures ou des motifs intéressants basés sur des critères prédéterminés, et extraire le plus de connaissances possible. Dans ce contexte, on a élaboré une étude de détections et prédictions de récidivité du cancer du sein.

a. La classification

C'est actuellement le mode d'apprentissage le plus couramment utilisé. Son principe est élémentaire : on soumet au classifieur un grand nombre d'exemples pour lesquels l'entrée et la sortie associée sont connues et les paramètres d'apprentissage sont modifiés de façon à corriger l'erreur commise par le classifieur (c'est-à-dire la différence entre la sortie désirée et la réponse du classifieur à l'entrée correspondante). Le classifieur a alors comme but d'approximer ces exemples aussi bien que possible et de développer à la fois la bonne représentation mathématique qui lui permet de généraliser ces exemples pour ensuite traiter des nouvelles situations (qui n'étaient pas présentes dans les exemples). [19]

Le classifieur reçoit directement en entrée les couples «entrée» / «sortie désiré » ou « paramètres descripteurs »/ « classes d'appartenance ».

b. La prédiction

Il est utilisé pour trouver la sortie numérique. Comme la classification, l'ensemble de données d'apprentissage contient les valeurs d'entrée et de sortie numériques correspondantes. Sur la base de l'ensemble de données d'apprentissage, l'algorithme dérive un modèle ou un prédicteur. Lorsque de nouvelles données sont fournies, le modèle doit trouver une sortie numérique. Contrairement à la classification, cette méthode n'a pas d'étiquette de classe. Le modèle prédit des fonctions avec des valeurs continues ou ordonnées. [19]

Dans notre travail, et dans le but d'aide au diagnostic du cancer de sein, on a utilisé deux outils de classification à savoir ; les Séparateurs à Vaste Marge (SVM) et K--plus proches voisins.

3. Classifieurs choisis pour la détection du cancer du sein :

3.1. KNN (k-nearest Neighbors):

La méthode des K--plus proches voisins fait partie des méthodes d'apprentissage basées sur les instances (en anglais « instance based Learning »).en effet ces méthodes stockent les données d'apprentissage et n'effectuent de généralisation que si de nouvelles données se présentent. La phase de généralisation consiste alors à calculer leur relation avec les données d'apprentissage déjà stockées pour en déduire leur classification. [20]

C'est une méthode non paramétrique qui ne nécessite aucune hypothèse sur les classes. L'idée est simple : elle consiste étant donné un point x représentant une forme à reconnaître ; à déterminer la classe de chacun des k points les plus proches (au sens d'une distance) de x parmi l'ensemble d'apprentissage.la décision est alors d'affecter x à la même classe que celle de son voisin le plus proche. [20]

3.1.1. Algorithme

- Paramètre : le nombre k de voisins
- Donnée : un échantillon de m exemples et leurs classes
  - La classe d'un exemple X est c(X)
- Entrée : un enregistrement Y
  - Déterminer les k plus proches exemples de Y en calculant les distances
  - Combiner les classes de ces k exemples en une classe c.
- Sortie : la classe de Y est c(Y)=c
  - Le choix de la distance est primordial au bon fonctionnement de la méthode.
  - Les distances les plus simples permettent d'obtenir des résultats satisfaisants.
- Propriétés de la distance :

$d(A,A)=0$, $d(A,B)= d(B,A)$; $d(A,B) \leq d(A,C) + d(B,C)$

- distance euclidienne

Soit $X = (x_1,..., x_n)$ et $Y = (y_1,..., y_n)$ deux exemples, la distance euclidienne entre X et Y est :

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Autres distances :

- Sommation

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)}$$

- Distance euclidienne pondérée

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}w_i(x_i-y_i)^2}$$

3.1.2. **Principe**

On cherche dans l' ensemble d'échantillons les k observations les plus proches à l'observation à classer et on procède par vote ou moyenne.

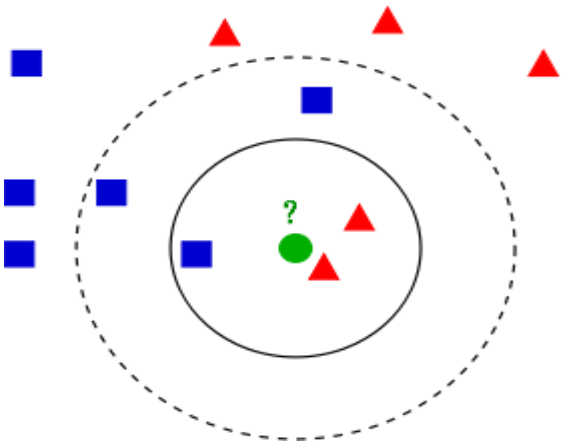


Figure 5: principe de KNN.

Les données : Il faut avoir un données d'apprentissage (Napp vecteurs de dimension d et étiquettes avec c classes) ainsi qu'un ensemble de test (Ntest vecteurs de dimension d et étiquettes). [20]

3.2. Les Support Vector Machines (SVM) :

Les Support Vector Machines souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination. [21]

Leur But est de trouver un classifieur linéaire (hyperplan) qui va séparer les données et maximiser la distance entre ces 2 classes.

Dans ce cas ces algorithmes sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation est la plus grande possible. [21]

Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation d'un modèle en contrôlant sa complexité. [21]

3.3. Objectif des SVM

L'objectif est de chercher Parmi les hyperplans valides, l'hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. Cette distance est appelée distance marge entre l'hyperplan et les exemples. [21]

Comme on cherche à maximiser cette marge, on parlera de méthode des séparateurs à vaste marge.

3.4. Algorithme :

Dans l'algorithme SVM, nous cherchons à maximiser la marge entre les points de données et l'hyperplan. La fonction de perte qui aide à maximiser la marge est la perte de charnière. La fonction de perte de charnière (la fonction de gauche peut être représentée comme une fonction de droite) et donnée par l'équation au-dessous[21]

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

Le coût est de 0 si la valeur prédite et la valeur réelle sont de même signe. S'ils ne le sont pas, nous calculons alors la valeur de la perte. Nous ajoutons également un paramètre de régularisation à la fonction de coût. L'objectif du paramètre de régularisation est d'équilibrer la maximisation de la marge et la perte. Après avoir ajouté le paramètre de régularisation, les fonctions de coût se présentent comme ci-dessous. [21]

Maintenant que nous avons la fonction de perte

$$min_w \lambda \| w \|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

,

Nous prenons des dérivées partielles par rapport aux poids pour trouver les gradients. En utilisant les gradients, nous pouvons mettre à jour nos poids par l'équation suivante [21] :

$$\frac{\delta}{\delta w_k}\lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k}\left(1 - y_i\langle x_i, w\rangle\right)_+ = \begin{cases} 0, & \text{if } y_i\langle x_i, w\rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Lorsqu'il n'y a pas d'erreur de classification, c'est-à-dire que notre modèle prédit correctement la classe de notre point de données, nous n'avons qu'à mettre à jour le gradient à partir du paramètre de régularisation. [21]

$$w = w - \alpha \cdot (2\lambda w)$$

Lorsqu'il y a une erreur de classification, c'est-à-dire que notre modèle se trompe sur la prédiction de la classe de notre point de données, nous incluons la perte avec le paramètre de régularisation pour effectuer la mise à jour du gradient.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

3.4.1. Evaluation de la performance du classifieur :

Dans la phase de test ont doit permettre l'affectation d'un nouvel objet à l'une des classes, au moyen d'une règle de décision intégrant les résultats de la phase d'apprentissage . L'objectif est d'obtenir une estimation la plus fidèle possible du comportement du classifieur dans des conditions réelles d'utilisation. Pour cela, des critères classiques comme les taux de classification et les taux d'erreur sont presque systématiquement utilisés. Mais d'autres critères, comme la spécificité et la sensibilité, apportent aussi des informations utiles.

a. Taux de classification

Les taux de classification et d'erreurs permettent d'évaluer la qualité du classifieur par rapport au problème pour lequel il a été conçu. Ces taux sont évalués grâce à une base de test qui contient des formes étiquetées par leur classe réelle d'appartenance comme celles utilisées pour l'apprentissage afin de pouvoir vérifier les réponses du classifieur. [21]

Le taux de classification correcte est défini par : $CC = \frac{vp(i) + vn(i)}{vp(i) + vn(i) + fp(i) + fn(i)}$

b. Sensibilité et spécificité

L'évaluation des performances d'un classifieur peut être réalisée par l'appréciation de deux lois statistiques, qui sont la sensibilité et la spécificité. [21]

La sensibilité Se(i) représente la probabilité de bonne classification de la classe i et la spécificité Sp(i) est une mesure indirecte de la probabilité de fausse alarme égale à 1- Sp(i).

Pour rappel, ces deux quantités sont définies par :

$$Se(i) = \frac{vp(i)}{vp(i) + fn(i)} \; Sp(i) = \frac{vn(i)}{vn(i) + fp(i)}$$

où les grandeurs V P(i), FN(i), V N(i), FP(i) sont définies dans le tableau suivant :

	Présence d'événement de classe i	Absence d'événement de classe i
Classification Positive	Vrai Positif VP(i)	Faux Positif FP(i)
Classification Négative	Faux Négatif FN(i)	Vrai Négatif VN(i)

Tableau 1: les définitions des grandeurs VP, VN, FP et FN.

4. Effet de la sélection des paramètres sur la performance de la classification :

La performance d'un classifieur dépend en grande partie de la qualité de la représentation des données à traiter, ce qui implique généralement l'obligation de représenter les données à l'aide d'un grand nombre d'attributs représentatifs. C'est souvent le cas lorsque certaines parties des données ne contiennent que des informations non pertinentes, redondantes ou inutiles pour la tâche de classification, ce qui complique cette dernière. Par conséquent, lors de la construction d'un système de classification, il est nécessaire de limiter le nombre d'attributs considérés afin d'optimiser ses performances. C'est exactement ce que fait le processus de "sélection des paramètres", qui vise à filtrer le vecteur d'attributs afin d'en extraire des informations distinctives et pertinentes qui améliorent la qualité de la classification. [21]

Dans notre cas on a utilisé Analyse composantes principal pour améliorer la performance de classifieur.

4.1. Analyse en composantes principal (ACP) :

L'analyse en composantes principales est une méthode de réduction du nombre de variables nécessaires pour représenter géométriquement les phénomènes. La réduction n'est possible que si les N variables initiales ne sont pas indépendantes.

L'ACP est une méthode dite factorielle, car la réduction ne consiste pas en une sélection des variables de base mais en une définition de nouvelles variables (principales), obtenues par combinaison des variables initiales. C'est une méthode linéaire.

5. Conclusion :

Dans ce chapitre, nous donnons un aperçu général du processus d'ECD et de ses différentes étapes, et l'exploration de données, les techniques utilisées (KNN, SVM) et la méthode de sélection des paramètres ACP afin de mieux comprendre ces outils pour nous aider à extraire des informations utiles et divers types de données. Dans le chapitre suivant, nous allons appliquer ces processus et interpréter les résultats obtenus.

Chapitre III

Expérimentations et Résultats

1. Introduction :

Dans la médecine les systèmes d'aide au diagnostic (classification et prédiction) ont un grand rôle dans la détection et la diagnostic des maladies afin de prévenir les maladies, aider les médecins et éventuellement sauver les individus.

Pour cela dans ce chapitre, on va proposer et discuter notre système qui est constitué de trois(03) principales étapes :

- Les définitions des bases de données et l'explication des paramètres de la BDD.
- La préparation des bases de données.
- L'interprétation des méthodes utilisées dans la classification et la prédiction.

2. Matériel utilisé :

	Spécifique	ASUS	SAMSUNG
Matériels informatique	Système d'exploitation	Windows 10 professional 64 bit	Windows 7 professional 64 bit
	Processeur	Intel® core™ i7-2630QM cpu @2.00GHz	Intel® core™ i5-3230M cpu @2.00GHz
	Stockage	500Go HDD	500Go HDD
	RAM	8GB	8GB (7.71Go utilisé)
Logiciel	Matlab 2014a.		
	Excel 2013.		

Tableau 2: Matériel utilisé.

3. Les bases de données utilisées

Nous avons utilisé deux(02) bases de données une pour la détection et l'autre pour la prédiction.

3.1. Base de données de la détection

3.1.1. La base de données WBCD entre 1989 -1991

Des échantillons arrivent périodiquement au fur et à mesure que le Dr Wolberg rapporte ses cas cliniques[22]. La base de données reflète donc ce regroupement chronologique des données. Ces informations de regroupement apparaissent immédiatement ci-dessous, après avoir été supprimées des données elles-mêmes[22] :

- Groupe 1:367 instances (janvier 1989).
- Groupe 2:70 instances (octobre 1989).
- Groupe 3:31 instances (février 1990).
- Groupe 4:17 instances (avril 1990).
- Groupe 5: 48 instances (août 1990).

- Groupe 6: 49 instances (mise à jour en janvier 1991).
- Groupe 7: 31 instances (juin 1991).
- Groupe 8: 86 instances (novembre 1991).
- Total : 699 points (à partir de la base de données donnée le 15 juillet 1992).

3.1.2. **Les attributs de la BDD :**

1. Numéro de code de l'échantillon : numéro d'identification
2. Épaisseur de l'agrégat (1 – 10) :
3. Uniformité de la taille des cellules : 1 – 10
4. Uniformité de la forme des cellules : 1 – 10
5. Adhésion marginale : 1 – 10
6. Taille des cellules épithéliales simples : 1 – 10
7. Noyaux nus : 1 – 10
8. Chromatine fade : 1 – 10
9. Nucléoles normales : 1 – 10
10. Mitoses : 1 – 11
11. Classe : (2 pour bénigne, 4 pour maligne).

a. Définition des attributs :

Paramètres	définitions
Épaisseur de l'agrégat	(1-10) Les cellules bénignes ont tendance à être regroupées en monocouches, tandis que les cellules cancéreuses sont souvent regroupées en multicouches.
Uniformité de la taille des cellules	(1-10) Les cellules cancéreuses ont tendance à varier en taille et en forme. C'est pourquoi ces paramètres sont précieux pour déterminer si les cellules sont cancéreuses ou non.
Uniformité de la forme des cellules	(1-10)Uniformité de la taille/forme des cellules : Les cellules cancéreuses ont tendance à varier en taille et en forme. C'est pourquoi ces paramètres sont précieux pour déterminer si les cellules sont cancéreuses ou non.

Adhésion marginale	(1-10) Les cellules normales ont tendance à se coller les unes aux autres. Les cellules cancéreuses ont tendance à perdre cette capacité. La perte d'adhérence est donc un signe de malignité.
Taille des cellules épithéliales simples	(1-10) Elle est liée à l'uniformité mentionnée ci-dessus. Les cellules épithéliales qui sont considérablement agrandies peuvent être des cellules malignes.
Noyaux nus	(1-10) Il s'agit d'un terme utilisé pour les noyaux non entourés de cytoplasme (le reste de la cellule). Ils sont généralement observés dans les tumeurs bénignes.
Chromatine fade	(1-10) Décrit une "texture" uniforme du noyau observée dans les cellules bénignes. Dans les cellules cancéreuses, la chromatine a tendance à être plus grossière.
Nucléoles normales	(1-10) Les nucléoles sont de petites structures visibles dans le noyau. Dans les cellules normales, le nucléole est généralement très petit, voire invisible. Dans les cellules cancéreuses, les nucléoles deviennent plus proéminents et sont parfois plus nombreux.
Mitoses	(1-10)Le cancer est essentiellement une maladie de la mitose incontrôlée.
Classe	(0 ou 1) Grosseur bénigne (non cancéreuse) ou maligne (cancéreuse) dans un sein.

Tableau 3: Définitions des paramètres.

f. L'ensemble de données du cancer du sein du Wisconsin.

Caractéristiques de l'ensemble de données :	Multivarié	Nombre d'instances :	699	Surface :	Vie
Caractéristiques des attributs :	Entier	**Nombre d'attributs :**	Dix 10	**Date du don**	1992-07-15
Tâches associées :	Classification	**Des valeurs manquantes ?**	Oui	**Nombre de visites Web :**	701617

Tableau 4 : Caractéristiques de l'ensemble de donnée, des attributset les tâches associées.

3.2. Base de données de la prédiction :

3.2.1. Informations sur l'ensemble de données :

C'est l'un des trois domaines fournis par l'Institut d'oncologie qui est apparu à plusieurs reprises dans la littérature sur l'apprentissage automatique. [23]

Cet ensemble de données comprend 201 instances d'une classe et 85 instances d'une autre classe. Les instances sont décrites par 9 attributs, dont certains sont linéaires et d'autres nominaux. [23]

3.2.2. Informations sur les attributs :

1. Classe : événements sans récidive, événements récurrents
2. Âge : 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90 -99.
3. ménopause : lt40, ge40, premeno.
4. taille de la tumeur : 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. nœuds inv. : 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps : oui, non.
7. deg-malig : 1, 2, 3.
8. Sein : gauche, droite.
9. poitrine-quad : gauche en haut, gauche en bas, droite en haut, droite en bas, central.
10. irradier : oui, non.

b. Définition des attributs :

Paramètres	Définitions
Âge	âge du patient au moment du diagnostic.
Ménopause	si la patiente est pré ou post-ménopausée au moment du diagnostic.
Taille de la tumeur	le plus grand diamètre (en mm) de la tumeur excisée.
Inv-nodes	le nombre (entre 0 et 39) de ganglions lymphatiques axillaires qui contiennent un cancer du sein métastatique visible à l'examen histologique.
node-caps	si le cancer métastase dans un ganglion lymphatique, bien qu'en dehors du site d'origine de la tumeur, il peut rester "contenu" par la capsule du ganglion lymphatique. Toutefois, au fil du temps, et en cas de maladie plus agressive, la tumeur peut remplacer le ganglion lymphatique puis pénétrer la capsule, ce qui lui permet d'envahir les tissus environnants.
Degré de malignité	le grade histologique (de 1 à 3) de la tumeur. Les tumeurs de grade 1 sont principalement constituées de cellules qui, bien que néoplasiques, conservent un grand nombre de leurs caractéristiques habituelles. Les tumeurs de grade 3 sont principalement constituées de cellules très anormales.
Sein	le cancer du sein peut évidemment se produire dans l'un ou l'autre sein ;
Quadrant mammaire	le sein peut être divisé en quatre quadrants, en utilisant le mamelon comme point central.
Irradiation	la radiothérapie est un traitement qui utilise des rayons X à haute énergie pour détruire les cellules cancéreuses.
Classes	(1) événements sans récidive (-1) événements avec récidive

Tableau 5: Définitions des paramètres.

g. Ensemble de données sur le cancer du sein :

Caractéristiques de l'ensemble de données :	Multivarié	Nombre d'instances :	286	Surface :	Vie
Caractéristiques des attributs :	Catégorique	**Nombre d'attributs :**	9	**Date du don**	1988-07-11
Tâches associées :	Classification	**Valeurs manquantes ?**	Oui	**Nombre de visites Web :**	560667

Tableau 6 : Caractéristiques de l'ensemble de donnée, des attributset les tâches associées .

4. Méthodes utilisées pour la classification et la prédiction :

Avant de passer aux méthodes de classification et prédiction, Le classifieur doit être évalué pour déterminer sa précision, son taux d'erreur et ses estimations d'erreurs. L'une des méthodes les plus primitives d'évaluation d'un classificateur est la "méthode d'exclusion". Dans cette méthode, l'ensemble de données est divisé de telle sorte que le maximum de données appartient à l'ensemble de formation et le reste à l'ensemble de test.

4.1. Méthode "Holdout" :

La méthode de retenue est le type le plus simple de validation croisée. L'ensemble de données est séparé en deux ensembles, appelés ensemble d'apprentissage et ensemble de test. L'approximateur de fonction ajuste une fonction en utilisant uniquement l'ensemble d'apprentissage. On demande ensuite à l'approximateur de fonction de prédire les valeurs de sortie pour les données de l'ensemble de test (il n'a jamais vu ces valeurs de sortie auparavant) [24]. Les erreurs qu'il commet sont accumulées comme précédemment pour donner l'erreur absolue moyenne de l'ensemble de test, qui est utilisée pour évaluer le modèle. L'avantage de cette méthode est qu'elle est généralement préférable à la méthode résiduelle et qu'elle ne prend pas plus de temps à calculer. Cependant, son évaluation peut avoir une variance élevée[24]. L'évaluation peut dépendre fortement des points de données qui se retrouvent dans l'ensemble d'apprentissage et ceux qui se retrouvent dans l'ensemble de test, et donc l'évaluation peut être significativement différente selon la façon dont la division est faite. [24]

```
[Training, Testing] = holdout (Data_Base_Name, Training_percentage);
```

4.2. Fonction "accuracy":

Le rôle de cette fonction est de calculer la précision en utilisant :

- TP : vrai positif,
- TN : vrai négatif,
- FP : faux positif,
- FN : faux négatif
- acc : précision (accuracy)

```
function [cm acc sens spes ] = Accuracy( y_actual,y_predicted )
```

5. Détection du cancer du sein :

5.1. Classification KNN :

5.1.1. Avant de l'analyse en composantes principales :

Dans cette étape nous avons utilisé le classifieur KNN sans méthode de sélection de variables ACP, Après plusieurs essais nous avons obtenus les résultats résumés dans le tableau 7.

Nous avons obtenu un taux de classification élevé (**99.42%)** alors qu'il faut voir en parallèle la Spécificité taux de réussite de l'algorithme KNN, de connaitre les cas normaux (**98.24%)** (Non malade) et le taux de sensibilité (**100%**) réussie de connaitre les cas anormaux (malade) en utilisant 75% des données pour l'apprentissage et 25% des données pour le test.

KNN	Données D'apprentissage	Taux de classification	Sensibilité	Spécificité
1	75%	97.14%	96.72%	98.11%
3	75%	**99.42%**	**100%**	**98.24%**
5	75%	98.28%	98.33%	98.18%
7	75%	98.28%	98.33%	98.18%
9	75%	98.28%	98.33%	98.18%
11	75%	98.28%	98.33%	98.18%

Tableau 7 : Les résultats obtenu utilisant la classification par KNN (K différents).

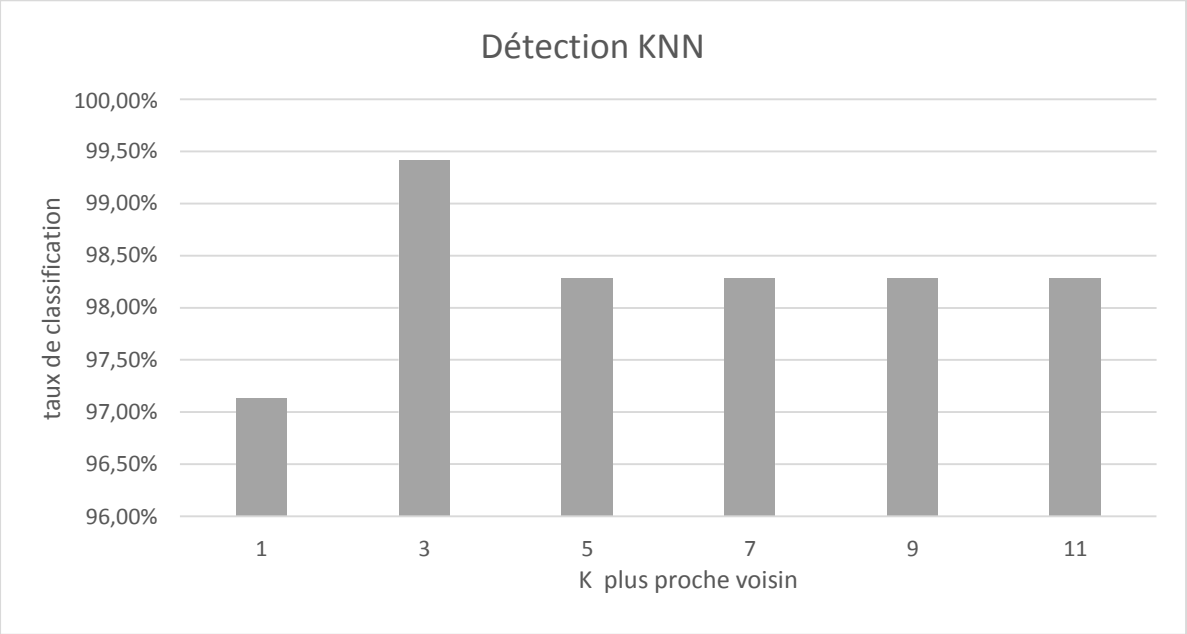


Figure 6 : Détection par KNN sans ACP.

5.1.2. Effet de sélection des paramètres sur la performance de la détection du cancer de sein :

Dans cette section nous allons appliquer la méthode ACP pour sélectionner les paramètres d'entrée du classifieur KNN, les résultats obtenus sont présentés dans le tableau 8.

KNN	Donnée D'apprentissage	**Taux de classification**	**Sensitivité**	**Spécificité**
1	75%	87.34%	86.56%	78.11%
3	75%	79.42%	81.42%	78.24%
5	75%	78.34%	81.37%	78.51%
7	75%	77.42%	81.28%	77.39%
9	75%	77.30%	81.17%	77.28%

Tableau 8 : Les résultats obtenus en utilisant la classification par KNN avec des K différents.

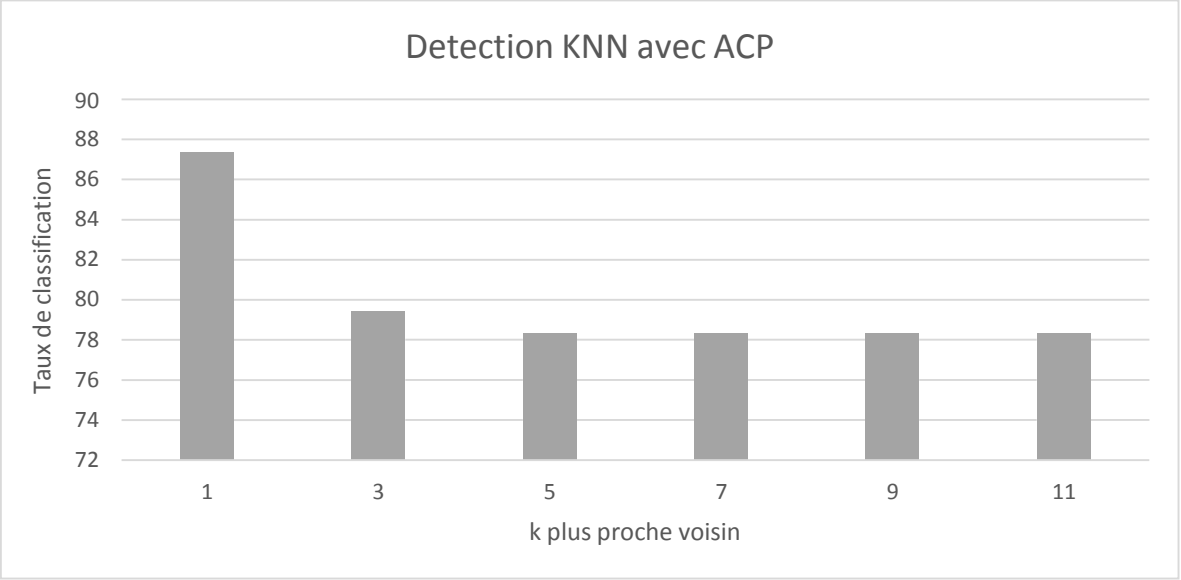


Figure 7 : Détection par KNN avec ACP.

Nous pouvons remarquer dans le tableau que le taux de classification diminue et cela peut être expliqué par lenombre limité des paramètres disponible (09).

Par contre l'usage de tous ces paramètres donne un résultat satisfaisant.

5.2. Classification par SVM :

Dans cette section nous allons utiliser un autre classifieur , les SVM pour la classification du cancer du sein, Après plusieurs essais les résultats obtenus sont résumés dans le tableau 9.Nous avons utilisé deux fonctions lors la prédiction en utilisant SVM (fitcsvm et svmtrain) afin d'avoir la meilleur fonction pour notre model.

a. fitcsvm

Cette fonction entraîne ou valide par croisement un modèle de machine à vecteurs de support (SVM) pour une classification à une ou deux classes (binaire) sur un ensemble de données prédicteurs de faible ou moyenne dimension. fitcsvm prend en charge le mappage des données prédicteurs à l'aide de fonctions noyau, et prend en charge l'optimisation minimale séquentielle (SMO), l'algorithme itératif à données uniques (ISDA), ou la minimisation de la marge douce L1 via la programmation quadratique pour la minimisation de la fonction objectif.

Mdl = fitcsvm(X,Y)

Renvoie un classificateur SVM entraîné en utilisant les prédicteurs dans la matrice X et les identifiants de classe dans le vecteur Y pour une classification à une ou deux classes.

b. svmtrain

Mdl = svmtrain(Training,Group)

- **Apprentissage**

Matrice de données de formation, où chaque ligne correspond à une observation ou un réplicat, et chaque colonne correspond à une caractéristique ou une variable. svmtrain traite les NaN ou les chaînes vides dans Training comme des valeurs manquantes et ignore les lignes correspondantes de Group.

- **Groupe**

Variable de regroupement, qui peut être un vecteur catégorique, numérique ou logique, un vecteur cellulaire de chaînes de caractères ou une matrice de caractères dont chaque ligne représente une étiquette de classe. Chaque élément de Group spécifie le groupe de la ligne correspondante de Training. Group doit diviser Training en deux groupes. Group a le même nombre d'éléments qu'il y a de lignes dans Training. svmtrain traite chaque NaN, chaîne vide, ou 'undefined' dans Group comme une valeur manquante, et ignore la ligne correspondante de Training.

Nous avons obtenu un taux de classification élevée (**98.29%)** alors qu'il faut voir en parallèle la Spécificité taux de réussite de l'algorithme SVM, de connaitre les cas normaux (**98.33%)**(Non malade) et le taux de sensibilité (**98.18%**) SVM a réussi à reconnaitre les cas anormaux (malade)**, tout ça** En utilisant 75% des données pour l'apprentissage et 25% des données pour le test.

Fonction SVM	Donnée D'apprentissage	Taux de classification	Sensitivité	Sensitivité
fitcsvm	**75%**	**98.29%**	**98.33%**	**98.18%**
svmtrain	75%	92.71%	92.24%	95.17%

Tableau 9: Les résultats obtenu utilisant la classification par SVM .

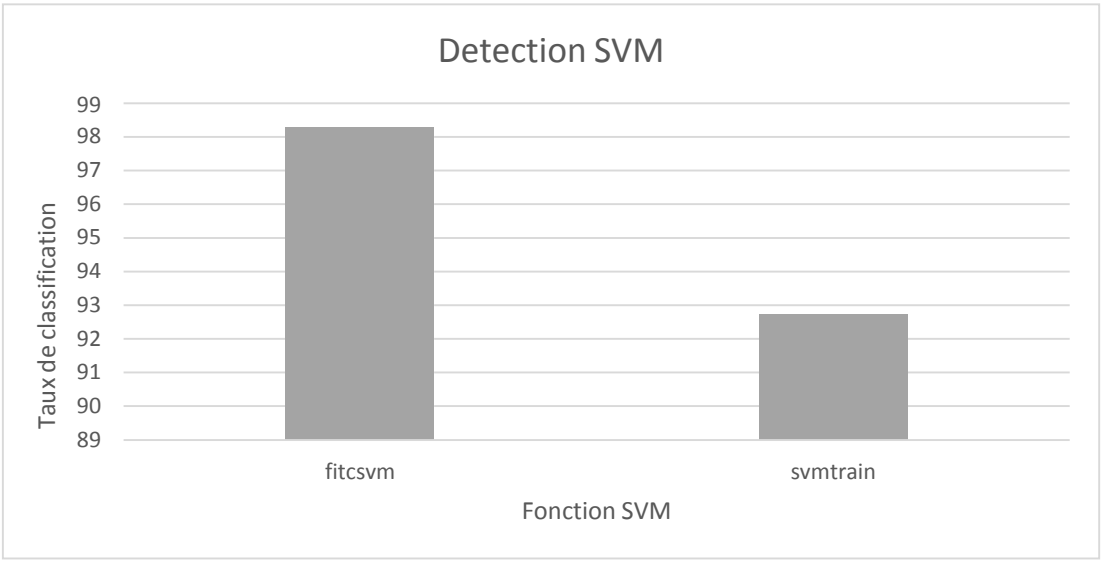


Figure 8 : Détection par SVM.

6. Comparaison

Après tous les essais réalisés nous remarquons que les résultats sont très proches .Cependant, si nous prenons en considération le temps de calcul, KNN représente le meilleur outil pour cette détection, car il n'a pas besoin d'une phase d'apprentissage.

Etant donné la complexité de la maladie, ou il faut minimiser au maximum le taux d'erreur donc le KNN représente le meilleur outil pour une détection optimale du cancer de sein.

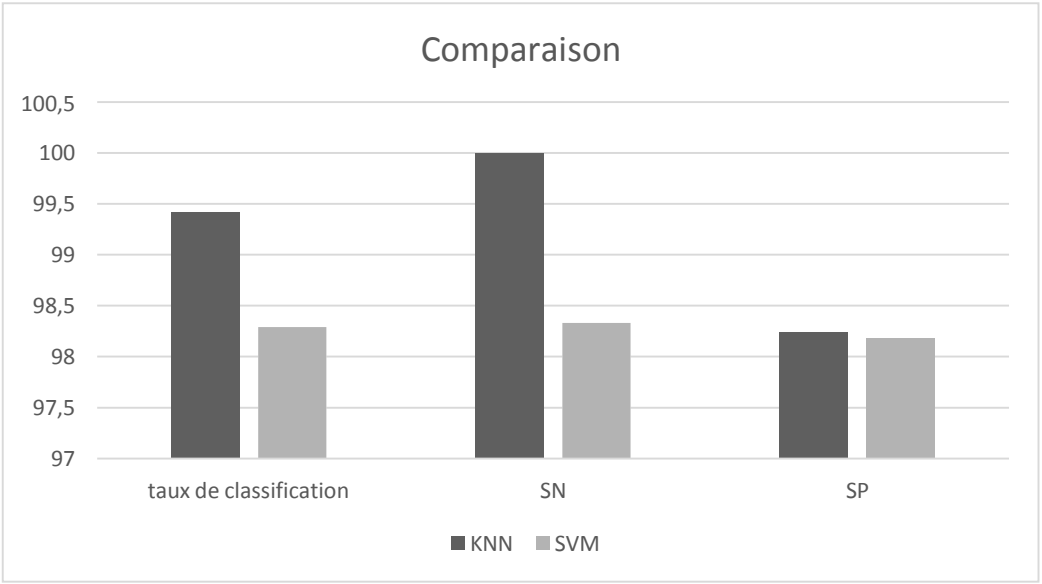


Figure 9 : Comparaison entre les Taux de classification entre KNN et SVM.

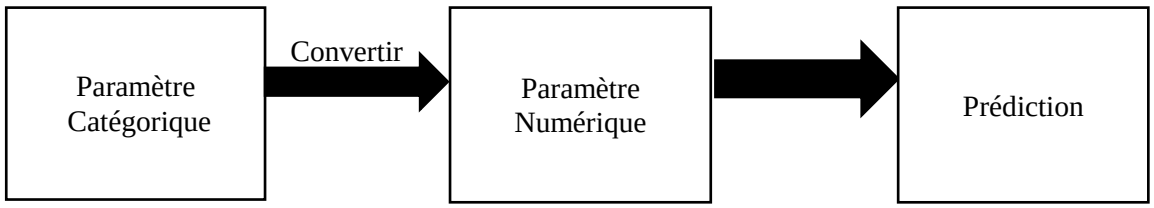
7. Prédiction de la récidivité :

7.1. Prétraitement de données (pre-processing) :

Après avoir présenté les résultats de la classification, nous allons passer à l'étape la plus importante qui est la prédiction du risque de récidive.

Nous allons utiliser une base de données différente pour la prédiction, avant cela nous devons d'abord prétraiter cette base pour pouvoir l'utiliser.

Dans la base de la récidive le type des paramètres est catégorique donc il faut passer par une phase de prétraitement des données avant de passer à la prédiction.



7.1.1. La conversion :

h. a. L'Age : nous calculons la moyenne d'âge.

Age	moyenne d'âge
10_19	14,5
20-29	24,5
30-39	34,5
40-49	44,5
50-59	54,5
60-69	64,5
70-79	74,5
80-89	84,5
90-99	94,25

Ménopause : nous avons utilisé une méthode appelée "one-hot-encoding".

Ménopause	lt40	ge40	premeno	codage_menopause
lt40	1	0	0	100
ge40	0	1	0	10
Premeno	0	0	1	1

i. Taille de tumeur : nous prenons le maximum.

Taille_Tumeur	Max_Taille
0-4	4
05_09	9
10_14	14
15-19	19
20-24	24
25-29	29
30-34	34
35-39	39
40-44	44
45-49	49
50-54	54
55-59	59

Univ-neud : nous prenons le maximum.

Univ-nœuds	Max_univ
0-2	2
03_05	5
6_8	8
9_11	11
12_14	14
15_17	17
18-20	20
21-23	23
24-26	26
27-29	29
30-32	32
33-35	35
36-39	39

j. Node-cap : pour oui (1), pour non (0).

Node_cap	nouveau_node
oui	1
non	0

Deg_malig : le type de ce paramètre est numérique donc il n y a pas besoins de conversion.

deg_malig
1
2
3

k. Sein : nous avons utilisé une méthode appelée "one-hot-encoding".

sein	gauch	droite	nouveau_P_Sein
gauche	1	0	10
droite	0	1	1

Quad-sein : nous avons utilisé une méthode appelée "one-hot-encoding".

quad-sein:	gauche-haut	gauche-bas	droite-haut	droite-bas	central	nouveau_P_quad_sein
gauche-haut	1	0	0	0	0	10000
gauche-bas	0	1	0	0	0	1000
droite-haut	0	0	1	0	0	100
droite-bas	0	0	0	1	0	10
Central	0	0	0	0	0	1

l. Irradiat : le type de ce paramètre est numérique donc il n y a pas besoin de conversion.

irradiat	N_P_Irradiat
oui	1
non	0

Classe : nous remplacent évènements sans récidive par (1) et évènements de récidive par (-1)

Labels	
événements sans récidive,	1
événements de récidive	-1

7.2. Prédiction par KNN :

Nous allons appliquer le KNN pour la prédiction de la récidive sur la base de données, nous avons choisis plusieurs K (nombre de voisins) et les résultats obtenus sont présentés dans le tableau 10. Nous avons obtenu un taux de classification (**96.43%)** alors qu'il faut voir en parallèle la Spécificité taux de réussite de l'algorithme KNN, de connaitre les cas normaux (Non récidive) (**100%)** dû à la présence d'un ensemble élevé des cas non-récidive par rapport aux cas récidive, et le taux de sensibilité (**96.30%)** réussie pour reconnaitre les cas anormaux (récidive).

Sachant que nous avons utilisé **90%** des données pour l'apprentissage et **10%** des données pour le test.

KNN	Données D'apprentissage	Taux de classification	Sensibilité	Spécificité
1	90%	78.57%	95.45%	16.66%
3	90%	82.14%	95.65%	20.00%
5	90%	85.71%	95.83%	25.00%
7	90%	85.71%	95.83%	25.00%
11	90%	82.14%	95.65%	20.00%
13	90%	89.28%	96.00%	33.33%
17	90%	85.92%	95.83%	25.00%
19	90%	89.28%	96.00%	33.33%
23	90%	92.85%	96.15%	50.00%
25	90%	**96.43%**	**96.30%**	**100%**

Tableau 10 : Taux de classification et sensibilité et spécificité utilisant KNN.

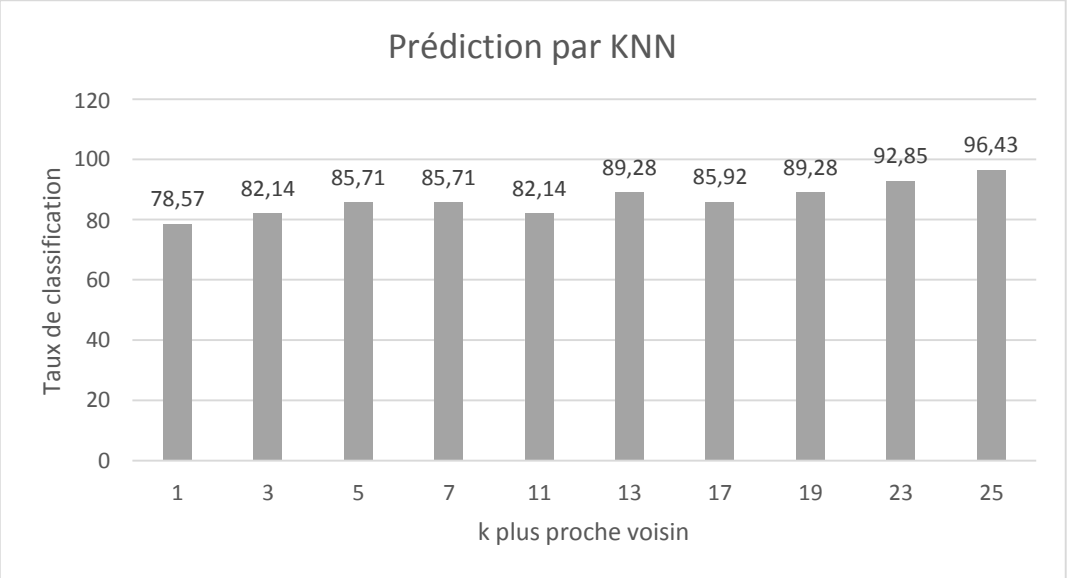


Figure 10 : Prédiction avec KNN.

7.3. Prédiction par SVM :

Nous allons présenter dans le tableau 11 les résultats de la prédiction en utilisant les SVM.

Nous avons obtenu un taux de classification (**78.57%)** alors qu'il faut voir en parallèle la Spécificité taux de réussite de l'algorithme SVM, de connaitre les cas normaux (**78.26%)**

(Non malade) et le taux de sensibilité (**80.00%**) réussie de reconnaitre les cas anormaux (malade)

En utilisant 90% des données pour l'apprentissage et 10% des données pour le test.

Fonction SVM	Données D'apprentissage	Taux de classification	Sensibilité	Spécificité
fitcsvm	**90%**	**78.57%**	**78.26%**	**80.00%**
svmtrain	90%	75.00%	76.92%	50.00%

Tableau 11 : Taux de classification et sensibilité et spécificité avec SVM.

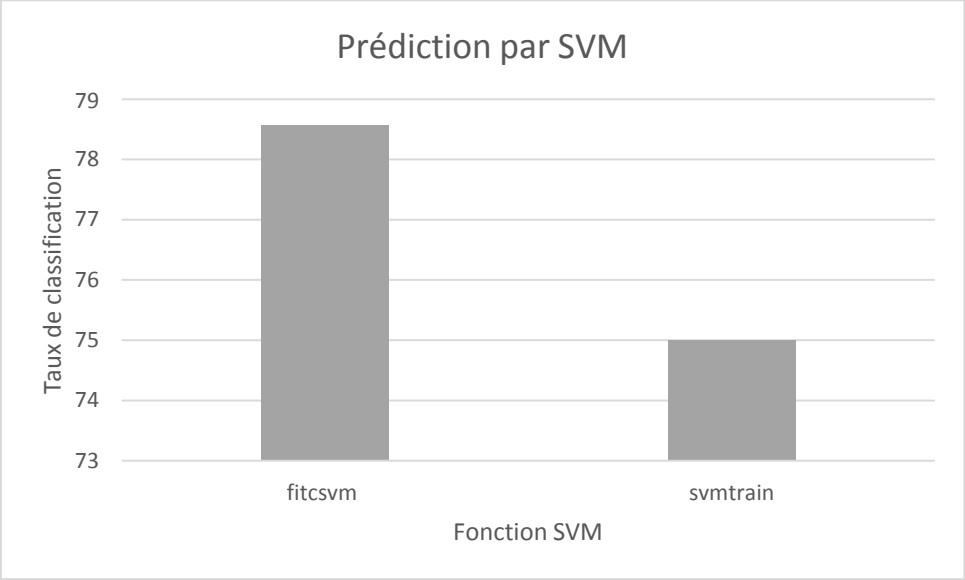


Figure 11 : Prédiction avec SVM.

7.4. Effet de sélection des paramètres sur la performance de la prédiction de la récidivité du cancer de sein :

Dans cette partie nous utilisons la méthode ACP pour sélectionner les paramètres d'entrée pour les KNN et SVM.

Dans les tableaux, nous pouvons remarquer que le taux de classification diminue et cela peut être expliqué par le nombre limité des paramètres disponible (09).

Par contre l'usage de tous ces paramètres donne un meilleur résultat.

KNN	Taux de classification	Sensibilité	Spécificité
1	64.28%	78.94%	33.33%
3	64.28%	78.94%	33.33%
5	64.28%	76.19%	28.57%
7	71.42%	78.26%	40.00%
11	60.71%	72.72%	16.66%
13	64.28%	73.91%	20.00%
17	60.71%	72.72%	16.66%
19	60.71%	72.72%	16.66%
23	64.28%	73.91%	20.00%
25	**75.00%**	**76.92%**	**50.00%**
27	**75.00%**	**76.92%**	**50.00%**

Tableau 12 : Taux de classification et sensibilité et spécificité utilisant KNN avec différents valeurs de K.

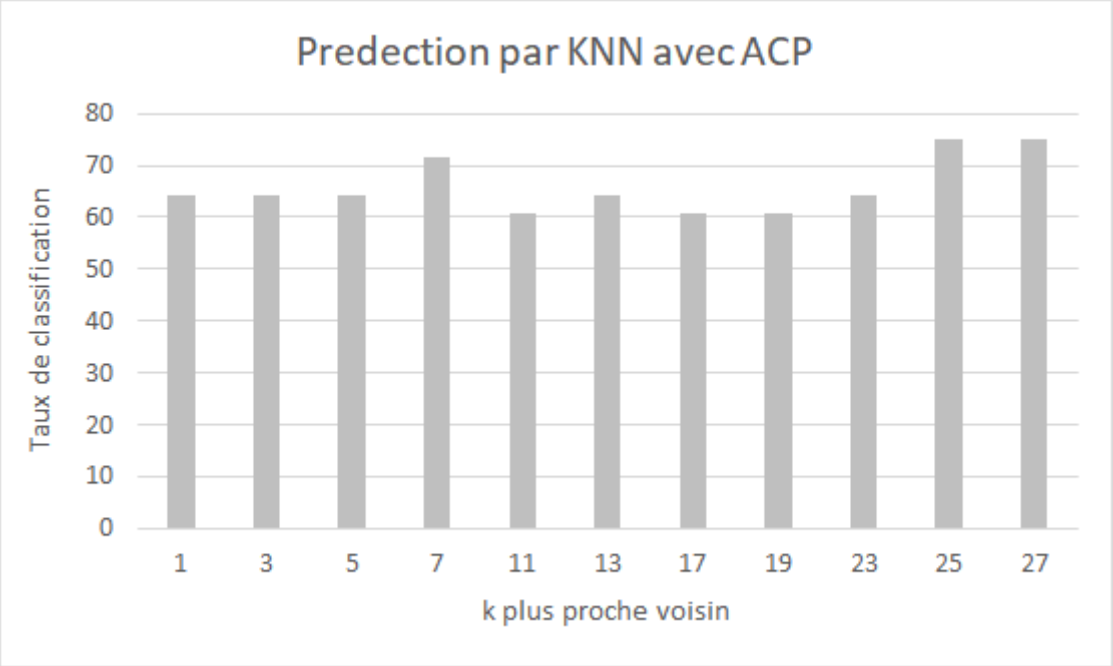


Figure 12 : Prédiction par KNN avec ACP.

Fonction SVM	Taux de classification	Sensibilité	Spécificité
fitcsvm	**78.57%**	**78.26%**	**80.00%**
svmtrain	75.00%	76.92%	50.00%

Tableau 13 : Taux de classification et sensibilité et spécificité en utilisant SVM.

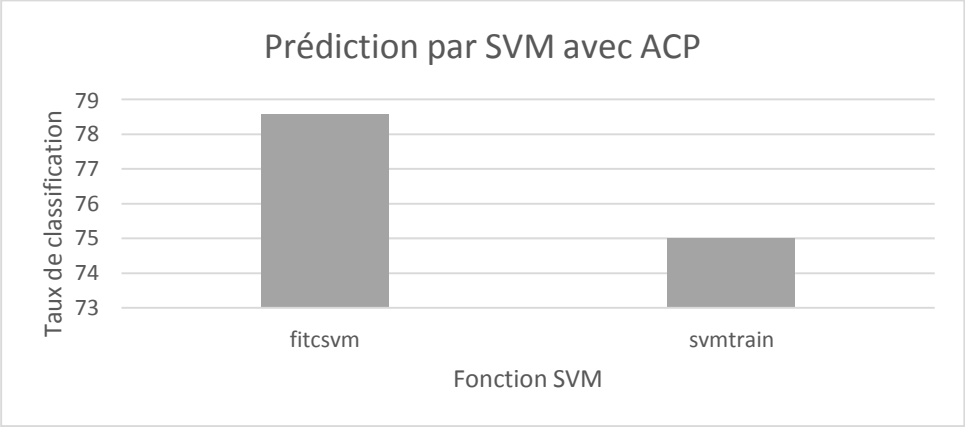


Figure 13 : Prédiction par SVM sans ACP.

7.5. Comparaison :

D’après les résultats présentés dans les sections, le KNN représente le meilleur résultat de prédiction de la récidivité du cancer de sein.

En prenant compte de la complexité de la maladie, où il faut minimiser au maximum le taux d'erreur donc le KNN représente le meilleur outil pour une prédiction optimale de la récidivité cancer de sein.

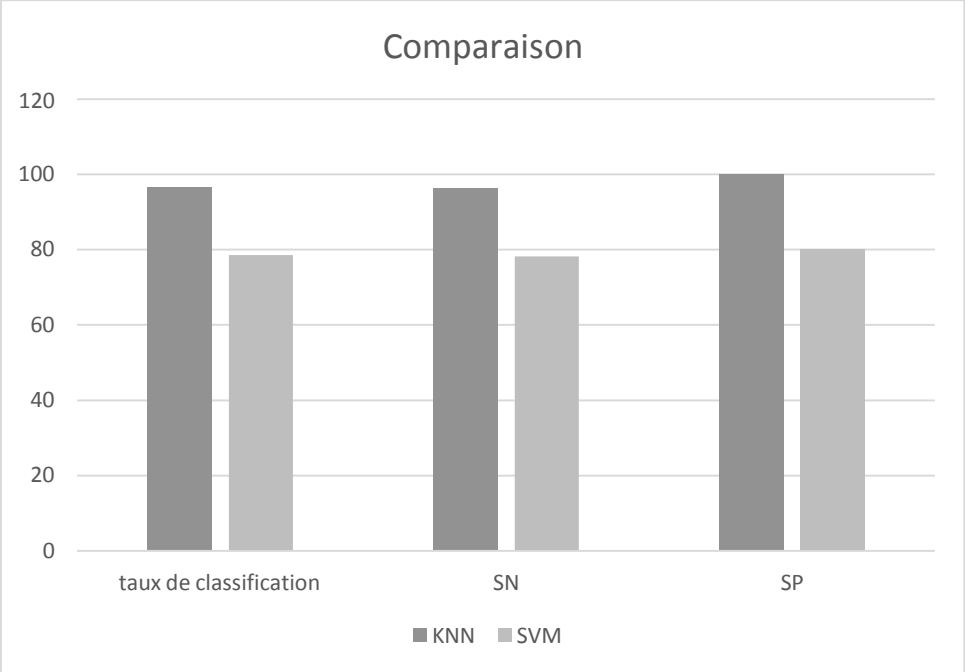


Figure 14 : Comparaison entre les résultats obtenus par SVM et KNN.

8. Comparaison avec les recherches précédentes :

Après avoir calculé les performances de notre méthode, nous avons comparé notre méthode avec les travaux de l’état de l’art dans la classification du cancer du sein. Nous présentons dans le tableau suivant les résultats en accuracy dans autres recherches en comparaison avec nos résultats.

Auteur	récidive		détection	
	Accuracy		Accuracy	
	KNN	SVM	KNN	SVM
Youness Khourdifi (2018)	96.1	83.52	96.1	97.9
Md. Milon Islam(2017)	95.68	77.43	97.19	98.57
Priyadh arshini R (2016)	95.42	80.12	97.08	97.09
Notre méthode	**96.43**	**78.57**	**99.42**	**98.29**

Tableau 14: La comparaison avec les recherches précédentes

La comparaison de notre méthode avec les autres existantes a montré que notre méthode a obtenu les meilleurs résultats surtout avec le classifieur KNN.

9. Conclusion

Dans ce chapitre, nous avons présenté la conception ainsi que la validation de notre approche. Nous avons commencé par la description de la conception générale de notre système suivie par une petite description des bases de données. Ensuite, nous sommes passés aux méthodes utilisées pour la détection du cancer du sein puis la prédiction de la récidive avec les classifieurs KNN et SVM. Enfin nous avons terminé notre chapitre par l'évaluation de notre système par les mesures de performances suivie par une comparaison de notre méthode avec les travaux de l'état de l'art.

Nous concluons par affirmation que notre système est un système robuste grâce aux résultats obtenus qui sont pertinents.

Conclusion générale

Dans ce mémoire de fin d'étude, nous avons brièvement présenté une recherche autours du cancer du sein pour réaliser un système d'extraction des données exploitables pour la classification de ce type de cancer et la prédiction du risque son récidivité, pour cela, premièrement nous sommes passés par un traitement et réorganisation de la base de données. Ensuite, nous avons fait une extraction des paramètres utilisant l'Analyse en Composantes Principales ACP, suivi par une classification par le K-plus proche voisins (KNN) et la Machine à Vecteur de Support (SVM) afin d'avoir une meilleure prédiction de récidivité et alors avoir la meilleure méthode compatible pour notre étude.

Le challenge principal de ce projet consiste à avoir la base de données car les données sur la récidive sont rarement enregistrées dans la plupart des ensembles de données sur le cancer du sein et de construire un modèle précis pour une meilleure prédiction de la récidivité du cancer du sein.

Ce travail peut toujours être développé du côté de la collecte de la base de données et du côté programmation (méthodes de classification et l'implantation de la machine Learning). Des recherches supplémentaires dans ce domaine devraient être menées pour améliorer les performances des techniques de classification afin de pouvoir prédire sur un plus grand nombre de variables.

Enfin, nous restons très optimistes quant à l'obtention dans un avenir proche d'un système qui répondra aux attentes des personnes et aider les médecins à suivre, surveiller et à détecter d'une manière précoce la récidivité d ce type de cancer.

Ce document est écrit dans l'intention d'être utilisé par de futurs chercheurs cherchant des informations liées à ce domaine.

Bibliographie

[1] M. F. Ak, « A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications », *Healthc. Basel Switz.*, vol. 8, n° 2, p. 111, avr. 2020, doi: 10.3390/healthcare8020111.

[2] « Breast cancer ». https://www.who.int/news-room/fact-sheets/detail/breast-cancer (consulté le juin 20, 2021).

[3] Z. Muhammad Zain, « Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis », *Int. J. Adv. Intell. Inform.*, vol. 6, p. 313‑ 327, nov. 2020, doi: 10.26555/ijain.v6i3.462.

[4] J. R. Harris, M. E. Lippman, U. Veronesi, et W. Willett, « Breast Cancer », *N. Engl. J. Med.*, vol. 327, n° 5, p. 319‑ 328, juill. 1992, doi: 10.1056/NEJM199207303270505.

[5] T. J. Key, P. K. Verkasalo, et E. Banks, « Epidemiology of breast cancer », *Lancet Oncol.*, vol. 2, n° 3, p. 133‑ 140, 2001, doi: https://doi.org/10.1016/S1470-2045(00)00254-0.

[6] « Anatomie du sein - Cancer du sein ». https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Anatomie-du-sein (consulté le mai 17, 2021).

[7] « InfoCancer - ARCAGY-GINECO - Cancer du sein - Formes de la maladie - La stadification ». http://www.arcagy.org/infocancer/localisations/cancers-feminins/cancer-du-sein/formes-de-la-maladie/la-stadification.html/ (consulté le mai 20, 2021).

[8] « Cancer du sein : dépistage, symptômes, pronostic, prise en charge ». https://sante.journaldesfemmes.fr/maladies/2505880-cancer-du-sein-symptomes-depistage-prise-en-charge-pronostic-guerison-survie/ (consulté le mai 17, 2021).

[9] Doctissimo, « Le cancer du sein en questions », *Doctissimo*. https://www.doctissimo.fr/html/dossiers/cancer_sein/9029-cancer-sein-questions.htm (consulté le mai 17, 2021).

[10] « Symptômes du cancer du sein - Société canadienne du cancer », *www.cancer.ca*. https://www.cancer.ca:443/fr-ca/cancer-information/cancer-type/breast/signs-and-symptoms/?region=on (consulté le mai 17, 2021).

[11] « Cancer du sein : la détection précoce - Dépistage du cancer du sein ». https://www.e-cancer.fr/Comprendre-prevenir-depister/Se-faire-depister/Depistage-du-cancer-du-sein/Cancer-du-sein-la-detection-precoce (consulté le mai 21, 2021).

[12] « Cancer du sein : dépistage, symptômes, pronostic, prise en charge ». https://sante.journaldesfemmes.fr/maladies/2505880-cancer-du-sein-symptomes-depistage-prise-en-charge-pronostic-guerison-survie/ (consulté le mai 21, 2021).

[13] « Dépistage du cancer du sein - le dépistage organisé ». https://www.roche.fr/fr/patients/info-patients-cancer/diagnostic-cancer/diagnostic-cancer-du-sein/depistage-organise-cancer-sein.html (consulté le mai 21, 2021).

[14] « Ponction cytologique - Diagnostic ». https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Diagnostic/Ponction-cytologique (consulté le mai 21, 2021).

[15] « Cancers du sein: le diagnostic | Fondation ARC pour la recherche sur le cancer ». https://www.fondation-arc.org/cancer/cancer-sein/diagnostic-cancer (consulté le mai 21, 2021).

[16] « Cancer : les traitements et les soins de support | Fondation ARC pour la recherche sur le cancer ». https://www.fondation-arc.org/traitements-soins-cancer/cancer-traitements-soins-de-support (consulté le mai 22, 2021).

[17] « Chirurgie (tumorectomie et mastectomie) - Cancer du sein ». https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Chirurgie-tumorectomie-et-mastectomie (consulté le mai 22, 2021).

[18] « Recurrent breast cancer - Symptoms and causes », *Mayo Clinic*. https://www.mayoclinic.org/diseases-conditions/recurrent-breast-cancer/symptoms-causes/syc-20377135 (consulté le mai 22, 2021).

[19] « [PDF] Extraction de connaissances à partir de données incomplètes et - Free Download PDF ». https://nanopdf.com/download/extraction-de-connaissances-a-partir-de-donnees-incompletes-et_pdf#modals (consulté le juin 25, 2021).

[20] I. José, « KNN (K-Nearest Neighbors) #1 », *Medium*, juin 02, 2021. https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d (consulté le juin 25, 2021).

[21] R. Gandhi, « Support Vector Machine — Introduction to Machine Learning Algorithms », *Medium*, juill. 05, 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (consulté le juin 25, 2021).

[22] « UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set ». https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29 (consulté le juin 25, 2021).

[23] « UCI Machine Learning Repository: Breast Cancer Data Set ». https://archive.ics.uci.edu/ml/datasets/breast+cancer (consulté le juin 25, 2021).

[24] « Cross Validation ». https://www.cs.cmu.edu/~schneide/tut5/node42.html (consulté le juin 25, 2021).