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 :DJEZZAR Meryem

Prédiction de la récurrence du cancer du sein par la Forêt de Corrélation Canonique

Soutenu le 26 août 2020 devant le Jury

Mme.	MEKKEOUI Nawel	*MAA*	Université de Tlemcen	Présidente
Mme.	SETTOUTI Nesma	*MCA*	Université de Tlemcen	Encadreur
M.	BECHAR Mohammed El Amine	*MCB*	Université de Tlemcen	Examinateur

Année universitaire 2019-2020

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABOU BEKR BELKAID
FACULTÉ DE TECHNOLOGIE
DÉPARTEMENT DE GÉNIE BIOMÉDICAL

MÉMOIRE DE FIN D'ÉTUDES

pour obtenir le grade de

MASTER EN GÉNIE BIOMÉDICAL

Spécialité : **Informatique Biomédicale**

présenté et soutenu publiquement
par

DJEZZAR Meryem

le 26 août 2020

Titre:

Prédiction de la récurrence du cancer du sein par la Forêt de Corrélation Canonique

Jury

Présidente du jury. Mme. MEKKEOUI Nawel, MAA UABB Tlemcen
Examinateur. Dr. BECHAR Mohammed El Amine, MCB UABB Tlemcen

Directrice de mémoire. Dr. SETTOUTI Nesma, MCA UABB Tlemcen

Année Universitaire 2019-2020

Je dédie ce mémoire à tous ceux qui m'ont épaulés et encouragés dans la réalisation de ce travail, à savoir :

A ma mère, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie et sa tolérance durant toutes mes années d'études. J'espère qu'Allah me donne la force et le courage pour que je puisse rendre ses sacrifices, pour toute son assistance et sa présence dans ma vie.

A mes sœurs, Sihem et Amel qui m'ont toujours soutenu et encouragé pour faire de mon mieux dans la vie, en plus de leur soutien et leur assistance, leur amour, leur aide et leur encouragement dans la réalisation de ce travail.

A mes proches amies plutôt ma deuxième famille : YAHIAOUI Marwa, Ben ABDELMALAK Marwa, MEKKI Marwa, BENAHCHILIF Souad, AINOUS Meriem, BOUKHOBZA Khadidja. Pour leur soutien permanent, leur patience et leur compréhension, leurs encouragements. Merci de savoir que je pourrais toujours compter sur vous et pour votre fidélité ainsi que le très grand appui moral et affectif tout le long de ces nombreuses années d'études. Un immense merci pour le bonheur que leur amour m'apporte.

A mes copines de promotion : FETTAH Abir, GOUMIDI Bouchra et tous mes amis de promotion pour les bons moments passées ensemble.

À toute ma famille.

À tous mes collègues de Génie Biomédical.

À toutes mes amies.

À tous ceux qui m'ont aidé de prés ou de loin.

Meryem

Remerciements

Je tiens tout d'abord à remercier ALLAH le tout puissant de m'avoir donné le courage, la force et l'aide nécessaire tout au long de mon parcours éducatif.

J'adresse mes sincères remerciements à toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques m'ont guidé et ont participé au succès de ce travail.

Je saisie cette occasion pour exprimer ma gratitude et mes vifs remerciements à mon encadreur, **Dr. SETTOUTI Nesma** pour avoir accepté de diriger ce travail, vous êtes une personne formidable, très attentive, pleine d'idées et toujours à l'écoute. Je vous remercie pour la qualité de votre encadrement, pour vos judicieux conseils, votre soutien, disponibilité, orientation et confiance. Bref, je vous remercie pour le bon déroulement de ce mémoire en cette période particulière.

Je présente mes sincères remerciements aux membres du jury : **Mme. MEKKEOUI Nawel** de nous honorer de présider ce jury, ainsi que **Dr. BECHAR Mohammed El Amine** d'avoir bien voulu faire partie de ce jury et examiner ce travail.

Mes remerciements s'adressent aussi à *Messieurs SAIDI Kamel, ACHIR MSIRDA Mohamed* dont les remarques pertinentes, les observations et les conseils m'ont permis d'améliorer ce travail.

Je remercie les différentes personnes, étudiants et enseignants, rencontrées au cours de mon cursus universitaire, au sein du département de génie biomédical.

J'adresse mes chaleureux remerciements à *ma mère*, ma source de force qui a été toujours là pour moi, pour l'encouragement moral et intellectuel le long de mes années d'études.

Je tiens à remercier, *Marwa*, une amie très chère à mes yeux qui a toujours cru en moi et qui m'a beaucoup aidé par son soutien inestimable et ses encouragements qui m'ont beaucoup aidé.

J'aimerais exprimer ma gratitude à tous ma famille, mes amis ainsi qu'à toute personne qui fera l'effort de lire ce document.

Résumé

La récidive est la réapparition d'un cancer à partir de cellules cancéreuses non détruites par le traitement initial. Ce phénomène ne ressurgit pas à une période bien précise, la possibilité de récurrence du cancer est alors une situation éprouvante à vivre, pour cela aujourd'hui, nombreuses sont les méthodes d'apprentissage automatique qui ont été appliquées pour améliorer les performances et augmenter l'efficacité des systèmes de prédiction de la récurrence du cancer. Durant cette dernière décennie, les chercheurs se sont plus intéressés aux méthodes d'ensemble tenant compte de leur précision, grâce à leur robustesse et capacité à préserver l'information de variabilité des données. Parmi elles : la forêt aléatoire, forêt rotationnelle et la forêt de corrélation canonique.

Dans notre projet de fin d'études, nous proposons d'utiliser les forêts de corrélation canoniques (CCF), une nouvelle méthode d'ensemble d'arbres de décision pour la prédiction de la récidive du cancer du sein. Cette méthode combine la puissance d'apprentissage des méthodes d'ensemble avec la force de discrimination de l'analyse canonique tout en augmentant la précision de chaque arbre et la diversité des arbres dans la forêt.

Les résultats expérimentaux appliqués sur la banque de données médicales récidive du cancer du sein montrent une amélioration de performance dans la tâche de la prédiction en comparaison avec la forêt aléatoire et la forêt rotationnelle en termes de taux de classification comme critère d'évaluation des performances.

Mots clés

La récurrence du cancer du sein, prédiction, méthodes d'ensemble, forêt aléatoire, forêt rotationnelle, forêts de corrélation canonique, analyse de corrélation canonique, UCI Machine Learning.

Abstract

Recurrence is the re-emergence of cancer from cancer cells not destroyed by the initial treatment. This phenomenon does not reappear at a specific period of time, the possibility of cancer recurrence is then a trying situation to live through, for this reason today, many automatic learning methods have been applied to improve the performance and increase the efficiency of cancer recurrence prediction systems. During the last decade, researchers have become more interested in ensemble methods taking into account their high precision, because of their robustness and ability to preserve the information of data variability. These include Random Forest, Rotational Forest and Canonical Correlation Forest.

In this Master Thesis, we propose to use Canonical Correlation Forests (CCF), a new decision tree ensemble method for predicting breast cancer recurrence. This method combines the learning power of ensemble methods with the discriminating power of canonical analysis while increasing the accuracy of individual trees and the diversity of trees in the forest.

Experimental results applied to the breast cancer recurrence dataset show an improvement in performance in the prediction task compared to Random Forest and Rotational Forest in terms of classification rate as a performance evaluation criterion.

Keywords

Breast cancer recurrence, prediction, ensemble methods, Random Forest, Rotational Forest, Canonical Correlation Forest, UCI Machine Learning Database.

ملخص

تكرار الإصابة بسرطان الثدي هو عودة السرطان بسبب ظهور الخلايا السرطانية التي لم يتم تدميرها بالعلاج الاول من جديد. هذه الظاهرة لا تظهر مرة أخرى في وقت محدد، وبالتالي فإن احتمالية تكرار الإصابة بالسرطان هي حالة صعبة للعيش، لهذا في يومنا هذا، تم تطبيق العديد من طرق التعلم الالي لتحسين الأداء وزيادة كفاءة أنظمة دعم التشخيص الطبي. خلال العقد الماضي، أصبح الباحثون أكثر اهتمامًا بأساليب المجموعات مع الأخذ بعين الاعتبار دقة نتائجهم نظرا لقوتهم وقدرتهم على الحفاظ على المعلومات وتنوع البيانات. من بينها: الغابة العشوائية، الغابة الدورانية، وغابة الارتباط القانوني.

في مشروع التخرج الخاص بنا، نقترح استخدام غابات الارتباط القانوني، وهي طريقة جديدة تابعة لأساليب مجموعات الاشجار التي تم تعيينها للتنبؤ بتكرار الإصابة بسرطان الثدي. تجمع هذه الطريقة بين القوة التعليمية لطرق التجميع والقوة التمييزية للتحليل القانوني مع زيادة دقة الأشجار الفردية وتنوع الأشجار في الغابة.

أظهرت النتائج التجريبية المطبقة على بنك البيانات الطبية (تكرار الإصابة بسرطان الثدي) أن غابات الارتباط القانوني أحسن أداءا لعملية التنبؤ مقارنة مع الغابة العشوائية والغابة الدورانية من حيث معدل التصنيف كمعيار للتقييم.

كلمات البحث

تكرار الإصابة بسرطان الثدي، التنبؤ، أساليب المجموعات ، الغابة العشوائية ، الغابة الدورانية ، غابة الارتباط القانوني، التحليل القانوني ، قواعد البيانات.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACC :Analyse de Corrélation Canonique.
ACP : Analyse en composantes principales.
CCF :Canonical Correlation Forest.
CCT :Canonical Correlation Trees.
CRF :Combination of Random Features.
DT :Decision Trees.
EMAPs :Extended Morphological Attribute Proles.
FN :Faux Négatif.
FP : Faux Positif.
LR : Logistic Regression.
LS-SVR :Least Squares Support Vector Regression.
MLC :maximum likelihood classier.
MRF :Markov random eld.
ODT :Oblique Decision Trees.
oob :out-of-bag.
PSO :Particle Swarms Optimization.
RCP : Réponse Complète Pathologique.
RF :Random Forest.
RFR :Random Forest Regression.
RL :Récidive Locale.
RLR :Récidives Locorégional.
RM :Récidive Métastatique.
RN :Réseau Neuronal.
Rot-For :Rotation Forest.
RR :Récidive Régionale..
RSS : Random Sub-Space.
SMOTE :Synthetic Minority Over-sampling Technique.
SVM :Support Vecteurs Machine.
TC :Taux de Classification.
UCI : University California Irvin.
VN : Vrai Négatif.
VP : Vrai Positif.
WPBC :Wisconsin Prognostic Breast Cancer.
XGBoost : eXtreme gradient boost.

Introduction

Le cancer du sein est un problème de santé majeur et l'une des principales causes de décès chez les femmes. La récidive survient lorsque le cancer réapparaît après quelques années de traitement. Pour faciliter le traitement médical et la prise en charge clinique, le diagnostic et le pronostic précoces du cancer sont devenus une nécessité pour éviter de la récidive du cancer du sein.

Comme les données médicales augmentent avec les progrès de la technologie médicale, l'exploration de données facilite la gestion des données et fournit une progression médicale et un traitement utile des conditions cancéreuses.

La prédiction précoce de la récidive du cancer du sein est l'un des travaux les plus cruciaux dans le processus de suivi. Diverses techniques d'apprentissage automatique peuvent être utilisées pour aider les médecins à prendre des décisions efficaces et précises et ainsi réduire le nombre de fausses décisions positives et négatives. Par conséquent, dans la littérature, diverses approches sont proposées pour arriver à une classification plus performante, parmi elles les méthodes d'ensemble afin améliorer et augmenter l'efficacité de classification et de prédiction.

Le principe de base de l'apprentissage d'ensemble consiste à construire un ensemble de classifieurs (c'est-à-dire à former plusieurs apprenants de base) et à classer ensuite les nouveaux échantillons en votant à la majorité de leurs prédictions. La précision des classifieurs individuels et la diversité dans l'ensemble sont des facteurs clés pour la performance globale des méthodes ensembles.

Dans ce projet de fin d'études, nous nous intéressons au domaine de la prédiction de récidive du cancer du sein, qui est devenu l'un des sujets de recherche les plus actif et récent dans le secteur de l'apprentissage automatique. Après une recherche bibliographique assez approfondie, nous avons choisis la dernière méthode d'ensemble en date à savoir, la forêt de corrélation canonique, qui a reçu une attention croissante pour nombreux problèmes de prédiction, se plaçant comme une nouvelle référence de performance dans le domaine.

Pour cela ce mémoire sera réparti comme suit :

Chapitre 1 : Contexte d'étude : La récidive du cancer du sein, nous avons présenté un aperçu général sur la récidive de cette maladie, un état de l'art sur les méthodes d'apprentissage automatique pour la prédiction de la récidive du cancer du sein, muni d'une synthèse de tous les travaux dans la littérature.

Chapitre 2 : Principe d'étude : La forêt de corrélation canonique, nous citons les points essentiels à la compréhension des principes de la forêt aléatoire, la forêt rotationnelle et la forêt de corrélation canonique, ainsi que les différentes applications de l'approche proposée.

Chapitre 3 : Application de la Forêt de Corrélation Canonique pour la prédiction de la récidive du cancer du sein, nous présentons le protocole d'expérimentation, aussi bien que les résultats expérimentaux obtenus et leurs interprétations.

En dernier lieu, une conclusion générale avec les perspectives futures envisagées.

Chapitre 1

Contexte d'étude : La récidive du cancer du sein

Introduction

Le cancer ou la tumeur maligne, est une maladie caractérisée par la présence des cellules anormales qui se proliférer au sein d'un tissu normal de l'organisme de telle manière que la survie de ce dernier est menacée. Dans le cas du cancer du sein, les cellules du sein deviennent incontrôlables, son type dépend des cellules du sein qui se transforment en cancer.

En Algérie, 44000 nouveaux cas de cancer, tous types confondus, sont annuellement enregistrés. En outre le cancer du sein est le premier en Algérie en terme d'incidence chez les femmes sachant qu'environ 12000 nouveaux cas sont enregistrés chaque année [1].

Le cancer peut réapparaître après le traitement au même endroit où il a commencé, c'est la récidive qui peut mettre en jeu le pronostic vital. Il existe plusieurs caractéristiques qui peuvent avertir qu'un cancer est en récidive, on parle alors des facteurs de risque.

D'après la littérature, différents algorithmes d'exploration de données tels que l'arbre de décision, support vecteur machine, réseau de neurone, sont explorés pour la prédiction de la récidive tenant compte de leur simplicité ainsi que de leur efficacité.

Dans ce chapitre, nous allons introduire la notion de la récidive, par la suite, il nous semble important de citer quelques travaux précurseurs concernant les techniques d'apprentissage automatique pour la prédiction des facteurs de risque de récidive du cancer du sein.

1 Le cancer du sein

Cette maladie est le résultat d'un dérèglement de certaines cellules responsable de la reproduction ce qui forme une masse appelée tumeur. Elle se développe à partir de la glande mammaire. Si les cellules cancéreuses appartiennent aux canaux galactophores, on parle de cancers canalaires et du cancers lobulaires si elles appartiennent aux lobules. Il existe aussi d'autres types de cancers du sein, beaucoup plus rares tels que les cancers médullaires, papillaires ou tubuleux.

Selon le stade d'évolution, différents types de cancer peuvent mise en place :cancer in situ ou intra canalaire (les cellules cancéreuses sont dans les canaux), cancer infiltrant ou invasif (les cellules cancéreuses sortent de la paroi des canaux). [2]

Pour le traitement du cancer du sein plusieurs méthodes sont utilisées tels que : la chirurgie, la radiothérapie, l'hormonothérapie, la chimiothérapie et les thérapies ciblées.

2 La récidive du cancer du sein

Une récidive n'est pas un second cancer mais la réapparition du premier cancer dans le même endroit que le cancer initial, ou dans un organe plus ou moins proche après une période de traitement complet.

La récidive du cancer du sein est comme toute récidive un choc violent pour le malade. La plupart des récidives du cancer du sein surviennent dans les 5 ans qui suit le traitement, mais ne sont pas systématiques car elles peuvent être beaucoup plus tardives et le malade doit donc continuer régulièrement les contrôles chez son médecin.

2.1 Les types de récidive du cancer du sein

En fonction de sa localisation, il existe trois types de récidive :

Récidive locale (RL) : lorsque le cancer du sein réapparaît dans la région de la poitrine ou du sein, ou dans la peau près du site d'origine ou de la cicatrice, (i.e. à la même partie du sein où il a été diagnostiqué plus tôt). [3]

Récidive régionale (RR) : lorsque les cellules cancéreuses du sein de la tumeur primaire se propagent aux ganglions lymphatiques (glandes) autour du sein. Ces derniers sont situés au niveau de l'aisselle, autour du sternum et entre les côtes (appelées ganglions mammaires internes), ou les ganglions au-dessus et en dessous de la clavicule. [3]

Récidive métastatique (à distance ou avancée) (RM) : se produit lorsque les cellules cancéreuses du sein de la tumeur primaire se propagent à d'autres parties du corps en passant par le système lymphatique ou sanguin. Les parties du corps où le cancer du sein se propage le plus souvent sont les os, le foie, les poumons

et le cerveau. Ce type de récidive peut être contrôlé pendant des années mais ne peut être guéri. [3]

Pendant le diagnostic, les deux premiers types de récidive sont interprétés comme des récidives locorégional (RLR) car ses symptômes sont similaires, et un autre diagnostique pour le troisième type. Les médecins cancérologues luttent contre le risque de récidives par des traitements succédant au traitement initial appelés traitements adjuvants.

2.2 Les facteurs de la récidive du cancer du sein

Cette maladie est multi-factorielle car elle est influencée par plusieurs facteurs morphologiques, pathologiques et biologiques des tumeurs pour la prédiction de la récidive dans différents intervalles. Les spécialistes se joignent a citer les facteurs suivants [4] :

- L'âge de la patiente au moment du diagnostic.
- La ménopause, si la patiente est pré- ou post-ménopausée au moment du diagnostic des antécédents familiaux.
- La taille de la tumeur : le pronostic est meilleur lorsque la tumeur est de petite taille.
- Le nombre de ganglions lymphatiques axillaires ou sus-claviculaire impliquées et enlevé : le risque augmente avec le nombre de ganglions atteints.
- La location de la tumeur : la région du sein dans laquelle se situe le cancer.
- Le grade histologique du cancer : varie de de I à III, plus le grade est élevé, plus les patientes sont à risque de récidive.
- Le statut des récepteurs hormonaux (RH) : dans le cas où tumeur possède des récepteurs à l'œstrogène et à la progestérone. Les tumeurs RH+ sont de bas grade car ils ont besoin d'une autre hormone pour se multiplier, donc moins agressives et moins susceptibles de se propager que les tumeurs dont les récepteurs hormonaux sont négatifs (RH-).
- Le type de la chirurgie ou la radiothérapie.
- La thérapie hormonale ou l'hormonothérapie : n'est utilisée que pour les cancers du sein qui ont des récepteurs pour les hormones. Elle est utilisée avant la chirurgie pour réduire la taille d'une tumeur et, après la chirurgie pour réduire le risque de récidive.
- La fraction de phase S : est une mesure du pourcentage de cellules dans une tumeur qui se trouvent dans la phase du cycle cellulaire au cours de laquelle l'ADN est synthétisé.

2.3 Les modalités de détection de la récidive du cancer

La détection de la récidive du cancer du sein implique des évaluations radiographiques, cliniques et autres. [5]

Évaluation clinique Effectuée à l'aide des trois tests utilisé pour la détection précoce dans le dépistage du cancer du sein comprennent la mammographie, l'examen clinique des seins et l'auto-examen des seins.

Évaluation en laboratoire De nombreux oncologues utilisent le sérum circulant marqueurs tumoraux, tels que CA 15-3 et CA 27-29, ou antigène carcinoembryonnaire (CEA) pour surveiller la récurrence de la maladie après la thérapie primaire.

Évaluation radiologique Les techniques de surveillance radiologiques jouent un rôle primordial pour la détection précoce d'une récidive de cancer du sein.

- Mammographie : est le pilier d'imagerie de la surveillance après le traitement curatif du cancer du sein. La surveillance régulière mammographique chez les femmes diagnostiquées à un stade précoce de cancer du sein améliore les résultats à long terme.

- Modalité supplémentaire : dans le cadre de la surveillance du cancer du sein. Nous parlons de l'IRM, de l'échographie et de la TEP/CT du sein. Sachant qu'on ne peut pas considérer l'un de ces derniers comme étant une technique de dépistage idéale. L'IRM offre une précision, une sensibilité et une spécificité élevées dans la détection des récidives locales et aussi dans la différenciation des cicatrices post-opératoire d'une tumeur récurrente.

- Il existe une autre technique d'imagerie importante qui s'intitule imagerie post-mastectomie qui consiste à identifier les algorithme et modalité de surveillance appropriés dans la population post-mastectomie afin de soutenir et décourager la routine de la surveillance par l'imagerie. Ainsi que d'autre études d'imagerie moléculaire, telles que les scanners osseux ou TEP-CT. Ces derniers ne sont pas indiqués sans plaintes ou constatations cliniques spécifiques sur l'examen physique contrairement à l'IRM ou les ultrasons .

3 L'apprentissage automatique pour la prédiction de la récidive du cancer du sein

Le modèle actuel de soins de suivi du cancer de sein est rétrospectif, long, coûteux et nécessite la réquisition de cliniciens, radiologues et chirurgiens, tout en sachant que les récidives sont détectées après la réapparition des symptômes, ce qui nous ramène à des résultats moins précis. Les techniques d'apprentissage automatique ont été mise en place dernièrement dans la prévention médicale [6], ces derniers permettent d'identifier le risque de récidive chez les femmes en se basant sur certains paramètres gynécologiques ainsi que d'autres caractérisant le cancer. Les programmes de surveillance qui utilisent cette approche favorisent l'amélioration des taux de survie tout en optimisant les traitements. A ce jour, plusieurs travaux (voir Santos et al. [7]) se sont intéressés à la prédiction de facteurs de risque de récidive du cancer du sein, dans cette section nous regroupant l'ensemble des propositions récentes.

- Aavula et al. (2019) [8] ont proposé un cadre extensible pour le pronostic du cancer du sein qui comprend la prédiction de susceptibilité, récurrence et survivabilité. Ils ont proposé un algorithme de sélection de sous-ensemble

de caractéristiques RSS (Random Sub-Space) appliqué au classifieur machine à support de vecteurs SVM pour améliorer l'efficacité du pronostic. Les performances de SVM-RSS sont comparées à différents algorithmes d'exploration de données : arbre de décision, SVM, RN (Réseau neuronal) pour la prédiction de ses trois aspects sur l'ensemble de données concernant l'incidence du cancer du sein et de la survie (SEER). Dans l'aspect récidive, l'algorithme proposé atteint les meilleurs résultats de prédiction.

– Mulatu & Gangarde (2017) [9] ont établi une étude comparative entre les différents algorithmes d'exploration de données avec un ensemble de données différents tel que : l'ensemble sur cancer du sein de l'Université du Wisconsin à partir de l'UCI, l'ensemble SEER concernant l'incidence du cancer du sein et de la survie, l'ensemble sur cancer du sein provenant de centre médical de l'Université de Leiden et l'ensemble de clinique de Kragujevac. Les résultats montrent que pour obtenir une valeur plus exacte sur la récurrence du cancer du sein, l'ensemble de donné de l'UCI est le plus adapté. De plus, les arbres de décision, support vecteurs machine et Bayes naïf permettent d'obtenir des résultats plus précis.

– Rasmussen et al. (2019) [10], ont réalisé une étude sur 471 femmes regroupées de quatre registres Danois (des registres de santé nationaux danois, le registre national danois des patients, le système d'enregistrement civil danois et le registre national danois de pathologie), afin de développer et valider un algorithme qui se base sur des critères d'exclusion limités pour identifier les patientes présentant une récidive du cancer du sein au Danemark. Cette approche offre une valeur de sensibilité de 97,3 % et de spécificité de 97,2 % et une valeur prédictive positive de 94,4%, aussi les dates de récurrence générées par l'algorithme ont montré un accord substantiel avec les dates de récurrence des 529 femme danoises opérées pour cancer du sein unilatérale à un stade précoce en 2003-2007.

– Mohebian et al. (2016) [11] ont proposé un système hybride de diagnostique assisté par ordinateur en utilisant les caractéristiques pathologiques et démographiques de 579 patientes atteintes d'un cancer du sein avec des informations complètes à Ispahan Centre de recherche sur le cancer Sayedo-Shohada, Iran. Ce système comparant la sélection des caractéristiques statistiques qui se devise en 3 méthodes : encapsulation, intégration et filtrage, L'optimisation des essaims de particules (PSO) pour filtrer les caractéristiques catégorielles et pour estimer les poids des entités d'intervalle, et la méthode d'ensemble « Forêt aléatoire » pour la prédiction de manière fiable et précise la récidive du cancer du sein au cours des 5 premières années après le diagnostic. Une comparaison en termes de performance avec les méthodes de réseau de neurones artificiels perceptron multicouches, support vecteur machine, arbre de décision, a permis de montrer que ce système offre des meilleurs résultats dont la sensibilité (77%), la spécificité (93%), la précision (85%).

– Asaoka et al. (2019) [12] ont effectué une analyse rétrospective des données pour 394 patientes qui ont atteint une réponse pathologique complète (RCP) au néo-adjuvant chimiothérapie. Les données ont été recueilli auprès de Centre médical de l'université de la ville de Yokohama, hôpital Yokohama Rosai, centre de cancer Kanagawa et Hôpital universitaire de médecine de Tokyo, Japon. Cette analyse est réalisée en utilisant la méthode Kaplan-Meier avec le test du logrank [1] pour l'analyse de la survie, le test exact de Fisher pour les comparaisons de groupe pour les variables catégorielles et le modèle de régression de Cox (modèle à risque proportionnel) pour l'évaluation des associations entre les résultats et les prédicteurs. Les résultats de cette étude sont parvenus au taux suivant : la récidive après l'obtention de la (RCP) est de 7.1%. Les taux de survie sans maladie à 5 ans (92,3%) et de survie globale (98,1%). Les principaux facteurs de la récurrence sont : le stade clinique avancé avant (NAC) déterminé par la taille de la tumeur, métastases ganglionnaires axillaires, et la positivité HER2.

– Chang et al. (2019) [13], ont utilisé 23 caractéristiques pour 2964 patientes diagnostiquées avec un cancer du sein dans trois hôpitaux : Hôpital universitaire médical de Chung Shan, Registre des tumeurs de l'hôpital de Jen-Ai et l'hôpital du Far Eastern, afin de développer un nouveau système de classification d''apprentissage automatique basé sur le classifieur boost de gradient eXtreme (XGBoost). Ce système vise à la prédiction des facteurs de risque de deuxièmes cancers primaires (DCP) chez les survivantes du cancer du sein, il se base sur 4 étapes :

  1. La transformation des données : pour une meilleure représentation de fonctionnalité par Analyse en composantes principales (ACP)),
  2. Le clustering : pour regrouper des cas similaires par l'algorithme k-moyenne.
  3. Le ré-échantillonnage : pour atténuer le déséquilibre des classes par la technique de sur-échantillonnage des minorités synthétiques.
  4. La technique d'apprentissage à travers le boosting de gradient extrême. Les techniques de surveillance radiologiques jouent un rôle primordial pour la détection précoce d'une récidive de cancer du sein.

  Les résultats montrent que le meilleur schéma est le XGBoost associé aux stratégies de ré-échantillonnage et de regroupement. Les facteurs de risque les plus importants associés avec des DCP chez les patientes atteintes d'un cancer du sein sont : l'âge, la séquence de la radiothérapie et de la chirurgie, marges du site primaire, facteur de croissance épidermique humain, cible clinique à dose élevée et les récepteurs d'œstrogènes.

– Ghasem Ahmad et al. (2013) [14], ont utilisé trois algorithmes d'apprentissage automatique (arbre de décision (ADD), réseau de neurones (RN), support vecteur machine (SVM)) pour développer un modèle de prédiction de

1. Le test du logrank est le test le plus populaire pour comparer plusieurs courbes de survie. C'est un test dit non-paramétrique. En effet, il permet de prendre en compte toute l'information sur l'ensemble du suivi sans la nécessité de faire des hypothèses sur la distribution des temps de survie.

récurrence de cancer du sein par l'analyse de 22 caractéristiques sur 1189 patientes du centre de cancer du sein à l'institut national de cancer, Tehran, Iran. Cette étude comparative montre que : les valeurs de précision sont de ADD (0.93), RN (0.94), et SVM (0.95) pour la prédiction du cancer du sein.

– Paredes-Aracil et al. (2018) [15] ont réalisés une étude sur 272 patientes atteintes d'un cancer du sein avec 10 variable initiales de diagnostic et 5 variable de thérapie néoadjuvante, afin de construire un modèle nommé COX, sous la forme d'un système de points, validé en interne par l'approche bootstrapping. La méthodologie statistique adoptée suit les points suivants : analyse fonctionnelle des prédicteurs continus, l'imputation multiple pour les données manquantes, la sélection de prédicteurs sur une base multi-factorielle. Les résultats montrent que :

  – Un pourcentage de 17.3% des patients ont développé une récidive en moyenne de temps de 8,6±3,5 ans.
  – Les variables importantes de ce système sont : l'âge, le grade, la multicentricité et le stade.
  – La validation par bootstrapping a montré une bonne discrimination et un bon calibrage.

  Ce modèle a été intégré dans une application mobile Android. « Breast Cancer Reccurence » pour prédire la récurrence du cancer du sein sur 5 à 10 ans.

A des fins de comparaison des résultats par la suite de ce projet de fin d'études, nous regroupons les travaux restant du domaine qui ont la particularité d'avoir utilisés la même base d'application dans le tableau ci-dessous :

AUTEUR	TITRE	MÉTHODE	APPLICATION	RÉSULTATS
Goyal et al. [16] 2020	*Prediction of Breast Cancer Recurrence : A Machine Learning Approach*	Ils ont fait un prétraitement des données : – Technique de sur-échantillonnage des minorités synthétiques – Élimination des données manquantes, – Normalisation des données – K-Means Clustering pour le type de cancer.	Ensemble de données (UCI) Machine Learning Repository, qui comprend 286 instances et 9 caractéristiques.	D'après la comparaison avec différentes méthodes (support vecteur machine, arbre de décision, Naive Bayes, réseau neuronal de régression généralisée et à propagation en avant), les meilleurs résultats sont obtenus pour les réseaux de neurones à propagation avant.

Ojha et al. [17] 2017	*A Study on Prediction of Breast Cancer Recurrence Using Data Mining Techniques*	Étude comparative entre les performances des algorithmes de classification et les algorithmes de clustering pour la prédiction.	Ensemble de données Wisconsin Prognostic Breast Cancer (WPBC) à partir D'UCI, il contient 190 instances et 35 attributs.	Les algorithmes de classification (plus précisément support vecteur machine, arbre de décision) sont meilleurs prédicteurs que les algorithmes de clustering (la plus précise est espérance-maximisation)
Pritom et al. [18] 2016	*Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique*	Ils ont fait un prétraitement des données : – Algorithme Ranker pour la sélection et Suppression des attributs redondants et non pertinents. – Application de différents algorithmes de classification à travers l'outil WEKA.	Ensemble de données Wisconsin Prognostic Breast Cancer (WPBC) à partir d'UCI, contient 190 instances et 35 attributs.	Ils ont déduit que : – Tout algorithme de classification est améliorée après Ranker (surtout l'arbre de décision et Naive Bayes) – Le support vecteur machine a fourni le meilleur taux de précision avec et sans application de l'algorithme Ranker en comparaison avec l'arbre de décision et Naive Bayes.
N. Alva [6] 2018	*Using machine learning techniques to predict the recurrence of breast cancer*	Ils ont utilisé un prétraitement : – Conversion de chaînes en valeurs numériques – Suppression des valeurs erronées. – Technique de sur-échantillonnage des minorités synthétiques – Application de différents clssifieurs.	Ensemble de données de cancer du sein de l'Institut d'oncologie, Ljubljana. Accessible dans le dépôt d'UCI, contient 286 instances et 10 variables.	Les valeurs de la précision données par cette étude sont : – Régression logistique :73,17% – Naive Bayes 73.17% – Support vecteur Machine :70% – Forêt aléatoire : 70.21% – K plus proche voisin : 67,54% – Arbre de décision : 64,27%

Dans un autre registre, actuellement, de plus en plus de techniques de prédiction de la récidive du cancer du sein se basent sur les approches biologiques. Ces méthodes sont centrées sur les paramètres clinique-pathologiques, les biomarqueurs, les scores des gènes et leurs relations. L'aspect de ses travaux reposent sur des analyses uni-variées de la récurrence de cancer du sein chez les femmes en utilisant des paramètres cliniques [19], en passant par, l'étude des relations entre paramètres cliniques, l'expression des gènes candidats et les score de récurrence des gènes [20].

Nous citons aussi dans un autre créneau, deux travaux concernent l'analyse de corrélation canonique qui analyse les relations entre les spécifications de la tumeur et les deux résultats à différents intervalles de temps pour construire un modèle de corrélation et qu'est basée sur les coefficients de structure au carré pour chaque variable pour déterminer les facteurs de risque importants de la récidive de cancer du sein, dont le pré-traitement de diffèrent ensemble de données est identique mais les résultat sont diffèrent.

- Sadoughi et al. [21], ont analysé les données relatives à la récidive du cancer du sein de 843 référées au Centre de recherche de cancer du sein (Téhéran, Iran) chez les femmes, ensuite ils ont sélectionné que 584 patientes (âge moyen de 45,9 ans) qui ont fait un suivi 5 ans après diagnostique comme échantillons d'étude. Le résultat abouti par cette étude est que les variables : antécédents familiaux, récepteur des œstrogènes, pathologie de la tumeur, type de chirurgie, taille de la tumeur et hormone thérapie sont des facteurs importants dans la prédiction de récidive locorégionale (LRR), entre et 5 ans ou plus.

- Razavi et al. [22] ont analysé les données relatives à la récidive du cancer du sein de 637 patientes cancéreuses du sein admises dans la région sud-est de la Suède. Ils ont trouvé que les variables : grade histologique de Nottingham, la fraction de phase S et nombre de ganglion lymphatique impliquées sont des facteurs importants dans la prédiction de récidive locorégionale (LRR) et Métastases à distance, dans les deux premières années.

4 Synthèse des travaux actuels

D'après les études précitées, nous constatons que plusieurs méthodes d'apprentissage automatique sont utilisées pour la prédiction de la récidive de cancer du sein telles que l'arbre de décision, support vecteur machine, réseau de neurones, ...etc. appliquées sur différentes bases de données comme UCI, SEER, centre de recherche de Iran, ...etc.

Chaque étude est réalisée à travers une étape de pré-traitement prédéfinie par son ensemble de données, comme le sur-échantillonnage des minorités synthétiques, la sélection des caractéristiques statistiques (RSS), l'optimisation des essaims de particules (PSO).... D'autre parts, il existe d'autre approche biologique qui associé des paramètres génétique, pathologique et cliniques. De même des

approches statistiques comme l'analyse canonique ont permis de trouver la relation entre deux ensembles de variable sur le même ensemble de données afin d'identifier les plus importants dans les deux ensembles des variables.

A travers ces constations, nous proposons une nouvelle approche qui exploite l'analyse canonique avec les méthodes d'apprentissage automatique pour la prédiction de la récidive du cancer du sein. Dans ce projet de fin d'études, nous étudierons une nouvelle méthode d'ensemble d'arbre de décision pour la classification et la régression, qui accepte plusieurs sorties, nommée Forêt de Corrélation Canonique (FCC) qui combine la puissance d'apprentissage des méthodes d'ensemble avec la force de discrimination de l'analyse canonique.

Conclusion

La récidive du cancer de sein est devenue une situation délicate et un phénomène constant pour les malades pour lesquels plusieurs travaux et recherches ont été mise en œuvre, en tenant compte que le processus d'investigation reste en développement continu.

Une grande attention a été consacré aux techniques d'apprentissage automatique afin d'établir un diagnostic médical à partir d'un ensemble de descripteurs cliniques d'un patient qui atteint une récidive.

Dans ce chapitre, nous avons recensé nombreux travaux, qui ont été mises en place afin de trouver des moyens fiables et plus développés pour détecter de manière précoce la récidive du cancer du sein d'où il est possible à présent de réaliser un diagnostic plus préalable.

Chapitre 2

Principe d'étude : La forêt de corrélation canonique

Introduction

Parmi les algorithmes d'apprentissage automatique, les méthodes d'ensemble tel que : le bagging [23], le boosting [24], les forêts aléatoires [25] ou plus récemment les forêts rotationnelles [26] et les forêts de corrélation canonique [27], ont gagné beaucoup en popularité ces dernières années dans la littérature.

Les méthodes d'ensemble sont des systèmes de classification multiple permettent de construire une collection de prédicteurs et agréger l'ensemble de leurs prédictions à partir d'un ensemble d'exemples représentatifs d'une population de données.

L'avantage principale de la combinaison de classifieurs est l'augmentation de la qualité, la précision, l'efficacité, la fiabilité et les performances des résultats par rapport à un système mono-classifieur.

Les méthodes d'ensemble d'arbres de décision telles que les forêts aléatoires (RF), forêts rotationnelles (Rot-For) sont largement utilisés pour les problèmes de classification et de régression en raison de la précision de chaque arbre individuel et la diversité de prédiction.

Dans ce chapitre, nous présentons la forêt de corrélation canonique, une méthode d'ensemble d'arbres de décision basée sur analyse de corrélation canonique. En premier lieu nous donnant un aperçu sur le principe des méthodes d'ensemble plus particulièrement ceux de la forêt aléatoire (RF-Random Forest) et la Forêt Rotationnelle (Rot-For-Rotate Forest) qui apporte une amélioration à la diversité de RF. En deuxième lieu, nous expliquons le principe de la forêt de corrélation canonique (CCF-Canonical Correlation Forest) avec qui apporte elle aussi une amélioration en utilisant l'analyse de corrélation canonique. A la fin, nous citons les travaux précurseurs concernant l'application de CCF dans différents domaines et nous terminons par une conclusion.

1 Principes de la forêt aléatoire et la forêt rotationnelle

La précision de chaque classifieur de base dans le modèle d'ensemble et la diversité entre les classifieurs représentent des facteurs clés pour évaluer les performances globales des ensembles de classifieurs.

Afin de mieux comprendre le principe de CCF, il nous semble important de rappeler quelque approche basique des méthodes d'ensemble. D'abord, la méthode de RF-Random Forest qui consiste en un groupe d'arbres de décision dont l'apprentissage de chaque arbre est basé sur un échantillon de bootstrap, puis la classification finale est effectuée sur la base d'un vote majoritaire des arbres dans la forêt. Aussi, la méthode de la Forêt Rotationnelle (Rot-For), qui peut être considérée comme une version améliorée du classifieur RF, elle vise à améliorer la diversité dans les arbres de décision constituant la forêt, en favorisant l'apprentissage de chaque arbre de décision sur l'ensemble des données dans un espace d'entité tourné qui rend le classifieur individuel aussi divers que possible.

1.1 La forêt aléatoire (RF)

La forêt aléatoire est une méthode statistique non paramétrique, la plus utilisée dans la littérature actuelle, appliquée à de nombreux problèmes de classification et de régression grâce à ces performances de prédiction, sa forte robustesse et le faible coût de temps. Introduit par Breiman [25], l'idée principale de l'algorithme comme décrit dans la Figure 2.1, se base sur l'agrégation d'une collection d'arbres de décision indépendants et distribués de manière identique sur des vecteurs aléatoires de même taille formé à partir des données d'entrée originales.

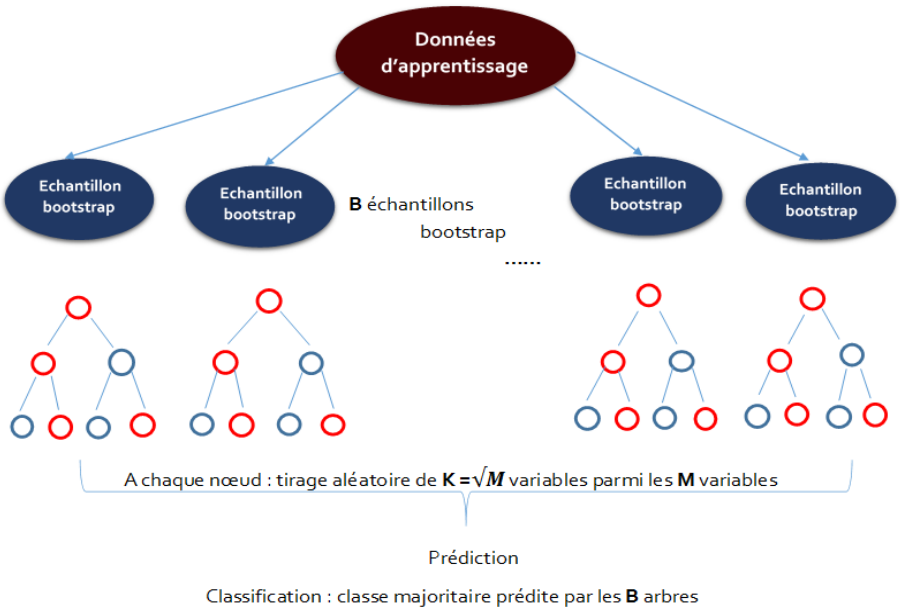

FIGURE 2.1 – Illustration schématique de la forêt aléatoire.

Environ 2/3 des échantillons sont utilisés pour former chaque arbre de décision en utilisant un échantillon bootstrap différent des données d'origine et le 1/3 restant laissés hors de l'échantillon, c'est oob (out-of-bag). Ces derniers sont utilisés pour obtenir une estimation d'erreur de classification ou de prédiction et pour mesurer l'importance des variables. La prédiction finale des forêts aléatoires est calculée par un vote majoritaire des prédictions individuelles des arbres en classification, ou la moyenne de ces derniers dans le cadre d'une régression.

Algorithm 1 Pseudo code de l'algorithme RF

Require: X l'ensemble d'apprentissage,
L le nombre d'arbres dans la forêt

1: **for** $i = 1$ **to** L **do**
2: $Bi \leftarrow Generer\ ensemble\ bootstrap\ de\ X$
3: $Ti \leftarrow Construire\ l'arbre\ ou\ chaque\ noeud$:
- Sélection aléatoire de $K = \sqrt{(M)}$ Variables à partir de l'ensemble d'attributs M.
- Identifier la variable la plus informative a partir de K représentant la valeur d'indice de Gini la plus grande.
- Création d'un nœud fils en utilisant cette variable.
4: $T \leftarrow T \cup Ti$
5: **end for**
6: **print** Ensemble d'arbres T qui composent la forêt.

1.2 La forêt rotationnelle (Rot-For)

La forêt rotationnelle ou Rotate Forest est une méthode d'ensemble d'arbres introduite par Rodriguez et al [26]. Elle est basée sur l'extraction de caractéristiques, en effectuant des transformations sur l'ensemble d'attributs par la division de façon aléatoire en K sous-ensembles (K fixé par l'utilisateur), par la suite, l'analyse des composantes principales (ACP) est appliquée à chaque sous-ensemble avant de construire chaque arbre.

L'analyse des composantes principales (ACP) [28] permet une représentation graphique des individus et des variables, afin de préserver les informations sur la variabilité des données et ainsi maintenir la précision individuelle et la diversité dans l'ensemble.

Une méthode de transformation linéaire est nécessaire pour projeter les données dans un nouvel espace de fonction pour chaque classifieur afin de tourner les axes ceci apporte une amélioration de l'exactitude individuelle en gardant tous les composantes et la diversité au sein de l'ensemble de la forêt par l'application d'extraction de caractéristiques.

En effet, d'après [26] [29], les forêts rotationnelles fournissent généralement des résultats meilleurs que d'autres classifieurs d'ensemble comme le Bagging, Boosting et les forêts aléatoires lorsque la taille de l'ensemble est relativement

petite. Il est à noter que Rot-For s'éloigne de RF de la Rot-For par le fait de transformer les attributs en ensembles de composantes principales par l'application de l'ACP, delà, tous les attributs sont utilisés pour chaque classifieur de base ces derniers sont de type l'arbre décision C4.5.

L'algorithme 2 reconstitue la démarche de la forêt rotationnelle.

Algorithm 2 Pseudo code de l'algorithme forêt rotationnelle

Require: X l'ensemble d'apprentissage.
Y la classe de la base d'apprentissage.
F l'ensemble de caractéristiques de la base.
L le nombre d'arbre.
n le nombre total d'attributs.
K le nombre de sous ensemble de caractéristiques.

1: **for** $i = 1$ **to** L **do**
2: Préparation de la matrice de Rotation R_i de taille, $n \times n$.
3: Diviser l'ensemble F en K sous-ensembles $F_{i,j}$ (avec $j = 1 \dots K$).
4: **for** $j = 1$ **to** K **do**
5: Soit $X_{i,j}$ l'ensemble de données associer à chaque sous ensemble.
6: Suppression aléatoire d'un sous ensemble de classe de $F_{i,j}$.
7: Sélection d'un échantillon Bootstrap de taille 75% de $X_{i,j}$ noté $X^{'}_{i,j}$.
8: Calcul des coefficients $C_{i,j}$ par l'application de l'ACP sur $X^{'}_{i,j}$.
9: Construction de la matrice de rotation arrangée R^a_i en coordonnant les composantes principales dans la matrice de sorte que chacun correspondent à la position de la variable dans le jeu de données d'apprentissage d'origine.
10: **end for**
11: Projetez l'ensemble de données d'apprentissage sur la matrice de rotation à l'aide de la multiplication matricielle.
12: Créez un arbre de décision T_i avec l'ensemble de données projeté.
13: **end for**
14: **print** Ensemble des arbres T qui composent la forêt.

2 La forêt de corrélation canonique (CCF)

La CCF est une nouvelle méthode d'ensemble d'arbre de décision proposée par Rainforth & Wood [27]. L'idée principale de l'algorithme comme décrite dans la Figure 2.2, est construite par plusieurs arbres de corrélation canonique individuels CCT (Canonical Correlation Tree) ou les arbres de décision oblique dont les divisions obliques sont des divisions d'hyperplan définies par une combinaison des caractéristiques, basés sur les coefficients de corrélation canoniques locaux calculés pendant l'apprentissage (section 2.2).

Avant d'expliquer et d'introduire la démarche de l'algorithme CCF, nous commençons par les principales différences entre les algorithmes d'apprentissage CCF et RF.

– Premièrement, pour les RF, chaque arbre est formé sur un échantillon de données bootstrap dans le bagging, mais pour les CCF, chaque arbre est formé sur l'ensemble des données d'apprentissage.
– Deuxièmement, dans la phase d'apprentissage de RF, un sous-ensemble aléatoire des caractéristiques est alors considéré à chaque nœud avec l'ensemble des individus correspond à ces caractéristiques. Cependant lors de l'apprentissage CCF, un sous-ensemble aléatoire des caractéristiques est également sélectionné, mais l'Analyse par Corrélation Canonique (CCA) avec projection de bootstrapping est utilisé en premier lieu pour projeter les caractéristiques dans l'espace des composantes canoniques, avec l'ensemble des individus divisés correspondant aux partitions uniques dans cet espace projeté.

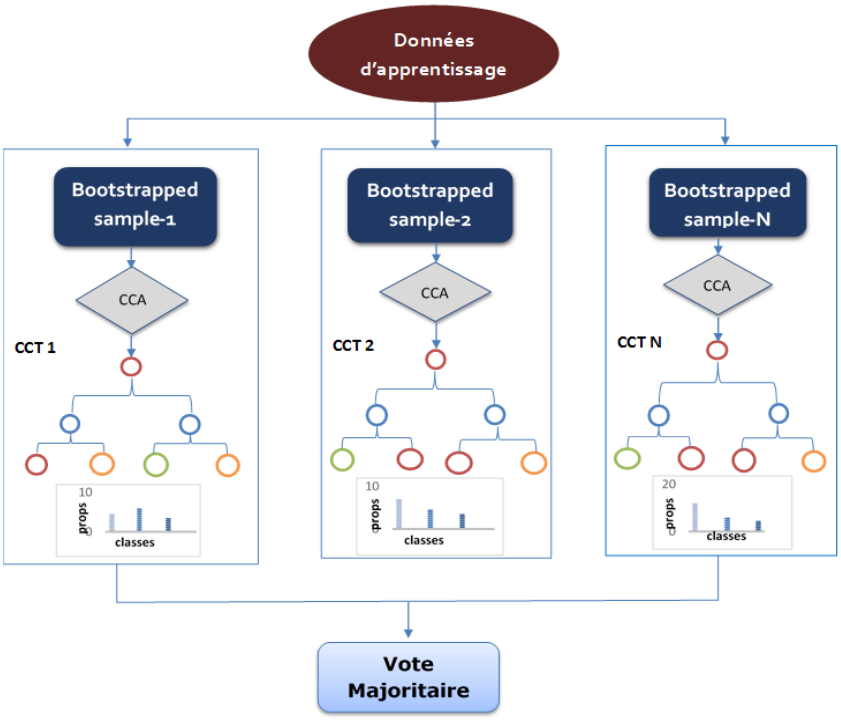

FIGURE 2.2 – Illustration schématique de la forêt de corrélation canonique (CCF).

2.1 L'arbre de corrélation canonique

Les arbres de corrélation canonique (Canonical Correlation Tree CCT) : sont des arbres de décision obliques, plus généraux aux hyperplans que les arbres de décision classiques. Ils produisent généralement de meilleurs résultats par rapport aux nombreux algorithmes, tels que la capacité de traiter efficacement les classes multiples, la stabilité numérique ou la diminution des coûts.

Une ODT (Oblique Decision Tree) [30] est une extension d'un arbre de décision classique avec des Combinaisons Linéaires. Cependant, ce type d'arbre en-

registre deux points de dissemblances importants :

1. Permet de trouver le meilleur hyperplan de séparation de l'ensemble d'apprentissage, ainsi n'impose aucune restriction sur son l'orientation,
2. Les hyperplans choisis sont obliques et ne sont pas nécessairement parallèles à l'un des axes.

L'apprentissage d'un CCT se fait en utilisant l'analyse de corrélation canonique (ACC) [31] (section 2.2) sur une projection bootstrap pour examiner la corrélation linéaire entre deux ensembles de variables et ainsi trouver des projections qui maximise la corrélation entre les entrées et sorties.

L'algorithme produit des arbres de décision petits, précis et ses calculs sont assez modestes, ce qui permet l'amélioration des performances.

2.2 L'analyse canonique

L'analyse de corrélation canonique (Canonical Correlation Analysis CCA) de Hotelling, 1936 [31] est une méthode déterministe pour calculer des paires de projections linéaires $\{A_\nu, B_\nu\}_{\nu=1:\nu\max}$ qui maximisent la corrélation entre deux matrices dans l'espace$\{WA_\nu, VB_\nu\}$.

Si nous appliquons l'ACC sur deux matrice arbitraire $W \in R^{n\times d}$ et $V \in R^{n\times k}$. La première paire de coefficients de corrélation sont donnés par :

$$\{A1, B1\} = argmax_{a\in R^d, b\in R^k}\ (corr(W_a, V_b)) \tag{2.1}$$

Les composantes de la corrélation canonique correspondante sont données par WA_1 et VB_1.

Les autres paires $\nu_{\max -1}$ sont créées où $\nu_{\max} = min(rang(W), rang(V))$ en répétant la même optimisation pour les nouvelles composantes qui ne sont pas corrélées avec toutes les composantes précédentes, par exemple :

$$(WA_1)^T WA_2 = 0 \ \ et \ \ (VB_1)^T VB_2 = 0 \tag{2.2}$$

Il est à noter que la solution de l'équation 2.1 peut être numériquement instable car elle nécessite une inversion de matrices de covariance typiquement dégénérées. Les chercheurs ont utilisé une combinaison des décompositions de QR (Q est matrice unitaire matrice, R est une matrice triangulaire supérieure) [32] et de SVD (Singular Value Decomposition) [33] pour trouver la solution de manière numériquement stable. Cette approche offre une plus grande stabilité numérique et une régularisation plus facile que les approches standard de calcul.

2.3 Principe de fonctionnement de la forêt de corrélation canonique (CCF)

La CCF est une approche d'apprentissage supervisé dont l'objectif est de faire des prédictions sur les classes à travers la séparation en hyperplan de vecteurs

correspondants aux caractéristiques d'entrée.

Comme pour les arbres de décision classiques, les CCT définissent une décomposition hiérarchique sur l'espace d'entrée, prend en considération qu'ont utilisé N entrées pour l'apprentissage $X = \{x_n\}_{n=1}^{N}$, avec leur classe $Y = \{y_n\}_{n=1}^{N}$.

Soit $T = \{Ti\}_{i=1...L}$ désigne une forêt canonique CCF, composée de L arbres de corrélation canoniques (CCT) individuels. Chaque arbre individuel $T_i = \{\Psi, \Theta\}$ est défini par un ensemble de nœuds discriminants $\Psi = \{\psi_j\}_{j \in J \setminus \partial J}$ et un ensemble de nœuds feuilles $\Theta = \{\theta_j\}_{j \in \partial J}$. Où J est un ensemble des indices des nœuds et $\partial J \subseteq J$ est un sous ensemble des indices des nœud feuilles.

Chaque nœud discriminant est défini par le n-uplet $\Psi_j = \{j, \delta_j, \phi_j, s_j, \chi_{j,r}, \chi_{j,l}\}$ où :
- j : l'identifiant unique du nœud,
- δ_j : un vecteur d'indices des caractéristiques utilisées pour le fractionnement au niveau du nœud,
- ϕ_j : un vecteur de poids utilisé pour projeter ces caractéristiques,
- s_j : le point auquel le fractionnement se produit,
- $\{\chi_{j,r}, \chi_{j,l}\}$ sont les deux identifiants de nœuds enfants.

La procédure de partitionnement est alors définie comme :

$$B(\chi_{j,l}, t) = B(j,t) \cap \{x \in R^D : x^T_{(\ \delta j)} \phi_j \leq s_j\}$$

$$B(\chi_{j,r}, t) = B(j,t) \cap \{x \in R^D : x^T_{(\ \delta j)} \phi_j \geq s_j\}$$

Avec $B(j,t)$: la partition de l'espace d'entrée associé au nœud j de l'arbre t. ($B(0,t) = R^D$ et $B(j,t) = B(\chi_{j,l}, t) \cup B(\chi_{j,r}, t)$

Chaque nœud de feuille est lui-même défini par :

$$\theta_j(x) = \frac{1}{N_j} \sum_{n \in \omega_j} y_i$$

Où :
ω_j : les indices des points dans le nœud j,donnée par $\omega_j = \{n \in 1...N : x_n \in B(j,t.)\}$.
N_j : le nombre correspondant de points d'entraînement au nœud.

Delà, la prédiction de la forêt est alors donnée par :

$$T(x) = \frac{1}{L} \sum_{i=0}^{L} Ti(x)$$

Les algorithmes 3 et 4 présentent le roulement étape par étape de l'apprentissage du CCF, en décrivant l'apprentissage de la CCF 3 et les processus de

croissance des arbres 4. L'apprentissage d'une CCF nécessite comme entrées les informations sur les données, le nombre d'arbres, nombre de caractéristique de sous-ensemble, la mesure d'impureté et le critère d'arrêts.

Algorithm 3 L'algorithme d'apprentissage CCF

Require: $\{X, Y\}$ les données,
L nombre d'arbre,
λ nombre de caractéristiques de sous ensemble $\lambda \in 1, \ldots, D$.
g mesure d'impureté
c critère d'arrêts

1: **for** $i = 1$ **to** L **do**
2: $T_i \leftarrow GROWTREE(0, X, Y, \lambda, g, c)$
3: **end for**
4: **print** Ensemble des arbres T qui composent la forêt.

Algorithm 4 L'algorithme de croissance d'arbre canonique

Require: identifiant du nœud racine j, x_j, y_j, λ, g, c.

- ID de sous-échantillon de caractéristiques δ_j de $\{1, \ldots, d\}\lambda$ fois sans remplacement .
- ensemble $\chi \leftarrow X^j_{(:,\delta\ j)}$
- construction d' un échantillon bootstrap de $\{\chi', Y^{j\prime}\}$
- calculer les coefficients de cca $\{\phi, \Omega\} \rightarrow cca\{\chi', Y'\}$
- projection des caractéristiques originales dans l'espace des composantes canoniques.
- choisissez la meilleure décomposition par leur gain (g)

1: **if** $g \leq 0$ **then**
Le nœud est une feuille, le modèle prédictif est la moyenne des points a la feuille.
2: **else**
3: Le nœud est un nœud discriminant.
4: Générer des identificateurs uniques pour les nœuds enfants $X_{j,l}$ et $X_{j,r}$
5: Attribuer des points de données aux nœuds enfants T^l et T^r
6: $\{\Psi_l, \theta_l\} \leftarrow GROWTREE(\chi_{j,l}, X^j_{(T^l,:)}, Y^j_{(T^l,:)}, \lambda, g, c)$
7: $\{\Psi_r, \theta_r\} \leftarrow GROWTREE(\chi_{j,r}, X^j_{(T^r,:)}, Y^j_{(T^r,:)}, \lambda, g, c)$
8: retour $\{\Psi_j \cup \Psi_l \cup \Psi_r, \theta_r \cup \theta_l\}$
9: **end if**
10: **print** Sous-arbre $\{\Psi_j, \theta_j\}$

Ces derniers ont des valeurs par défaut, cela démontre que les CCF est une méthode non paramétrique. Chaque arbre d'une CCF est formé de manière indépendante en utilisant l'ensemble des données. Le processus d'apprentissage de l'arbre (GROWTREE) repose sur les étapes principale suivants :

1. Choix du nœud racine avec l'ensemble des données.
2. À chaque itération, sélectionner une division optimale pour un ensemble de candidats générés.

3. Tester si le nœud courant est un nœud de feuille ou à un nœud discriminant.
4. Si le nœud est un nœud discriminant, appeler récursivement GROWTREE à nouveau pour produire des sous-arborescences pour chacun des nœuds enfants nouvellement générés.

Le processus d'apprentissage est terminé lorsque toutes les branches générées ont terminé en tant que nœuds de feuilles, le nœud est affecté à une feuille si aucune division n'est bénéfique (selon le gain de division) ou si un critère d'arrêt est rempli. A savoir les deux principaux critères d'arrêt possibles, sont la profondeur maximale de l'arbre, au-delà de laquelle tous les nœuds sont assignés comme feuilles (1 par défaut), et un nombre minimum de points de données contenus pour lesquels un nœud est autorisé à se diviser (par défaut est 2 pour la classification et 6 pour la régression).

3 État de l'art relatif aux forêts de corrélation canonique CCF

Au cours de ces dernières années, la forêt de corrélation canonique est devenue une approche d'actualité, qui présente un intérêt majeur pour diverses applications citées par la suite :

Rainforth & Wood [27] sont les initiateurs de cette nouvelle méthode d'ensemble pour la classification, la régression et la prédiction de sorties multiples. Nommée forêt de corrélation canonique, elle se base sur l'analyse de corrélation canonique dans l'apprentissage de chaque arbre individuel. Dans [27], ils effectuent plusieurs comparaisons des performances de CCF avec les méthodes d'ensemble : les forêts aléatoires (RF) et les forêts rotationnelles (Rot-For), ainsi qu'avec les 179 classifieurs considérés dans une enquête récente [34]. Les résultats prouvent que cette approche surpasse tous les classifieurs en termes de la classification.

Ha et al. [35], ont réalisé une étude comparative de trois algorithmes d'apprentissage automatique : forêts aléatoires (RF), les forêts rotationnelles (Rot-For) et les forêts de corrélation canonique (CCF) en comparaison à l'approche plus traditionnelle MLC classifieur de maximisation de probabilité pour cartographier la distribution en surface des communautés d'herbiers marins à faible et à forte couverture dans le port de Tauranga, New-Zélande. Cette étude indique que les techniques d'apprentissage automatique étaient plus performantes que le MLC avec la rotation (Rot-For) comme meilleur interprète, aussi c'est une approche efficace et prometteuse pour améliorer la précision de la surveillance des herbiers marins.

Dans un autre registre, Colkesen et al. [36], ont appliqué l'algorithme de la forêt de corrélation canonique (CCF) sur les images sentinelles pour la classification de l'utilisation des terres et de l'occupation des sols. Ils ont comparé les performances de CCF à celles des méthodes d'ensemble RF et Rot-For. Les principales déductions sont :

- Les algorithmes CCF et Rot-For produisent des résultats statistiquement similaires, mais ils ont tous deux surpassé l'algorithme RF.
- L'algorithme CCF a donné les meilleurs résultats pour les cas comportant un nombre inférieur d'échantillons, aussi, il a été jugé moins sensible à la taille de l'ensemble (c'est-à-dire le nombre d'arbres) par rapport à l'algorithme RF.

Xia et al. [37], ont proposé d'utiliser les forêts de corrélation canonique (CCF) pour la classification des images hyperspectrales sur six ensembles de données en intégrant trois stratégies pour étendre la CCF aux informations spectrales-spatiales : Champs aléatoires de Markov (MRF), profils multi-attributs étendus (EMAP), et l'ensemble de l'analyse des composantes indépendantes par filtre de guidage roulant (E-ICA-RGF). L'efficacité de cette approche a été évaluée en termes de précision et complexité de calcul, par rapport aux autres méthodes d'ensemble (les forêts aléatoire (RF) et forêts rotationnelle (Rot-For)), les résultats ont montré que le nouvel algorithme est une approche prometteuse et surpasse les autres algorithmes pour la classification des images hyperspectrales, mais aussi pour les informations spectrales-spatiales.

Dans un contexte de prédiction des données, Wang et al. [38] ont proposé un algorithme basé sur la forêt de corrélation canonique (CCF) avec une combinaison linéaire de caractéristiques, nommée the canonical correlation forest algorithm with a combination of random features (CCF-CRF) pour faire la prédiction du niveau des eaux souterraines à court terme, en utilisant les niveaux des eaux souterraines et les données météorologiques pour le champ de source de la rivière Daguhe en Qingdao, China. Pour évaluer l'efficacité du CCF-CRF une étude comparative avec la régression aléatoire améliorée des forêts (random forest regression (RFR)), et les moindres carrés pour la régression des vecteurs de support (least squares support vector regression (LS-SVR)) a été réalisée. Les résultats montrent que l'algorithme CCF-CRF proposé fournit des prédictions de niveau des eaux souterraines à court terme les plus précises, et a démontré qu'il offre un meilleur compromis entre la performance des prédictions et le temps de calcul en comparaison aux deux autres algorithmes.

Récemment, Ocansey et al. [39], ont développé un nouveau prédicteur de phosphosite basé sur la forêt de corrélation canonique (CCF), appelée CCF-Phos pour la prédiction des sites de phosphorylation dans les protéines de mammifères à partir de séquence d'acide aminé primaire. Ils ont évalué les performances de CCF-Phos avec d'autres méthodes populaires de prédiction des phosphosites existants chez les mammifères. CCF-Phos a obtenu globalement de bons résultats dans la sensibilité et la spécificité.

Dans une autre étude comparative, Sahin et al. [40] ont proposé d'utiliser la forêt de corrélation canonique (CCF) dans la prédiction de la sensibilité aux glissements de terrain pour le district de Yenice de Karabuk en Turquie. Ils ont testé la robustesse et la pertinence de CCF par rapport aux algorithmes des forêts aléatoires (RF), les forêts de rotation (Rot-For), et la régression logistique (LR). Les

résultats montrent que les méthodes d'ensemble étaient plus performantes que la méthode LR et robustes dans la prédiction spatiale des zones sensibles aux glissements de terrain. Ils ont conclu que CCF est un apprenant alternatif efficace pour produire des cartes de sensibilité précises, produisant ainsi des résultats statistiquement similaires aux algorithmes RF et Rot-For.

Étant donné ces derniers travaux de prédiction qui ont montré la capacité et le potentiel de prédiction des CCF, nous nous intéressons dans ce projet de fin d'études à l'application des CCF pour la prédiction de la récidive du cancer du sein comme une des techniques de génération d'ensemble d'arbre indépendant la plus performante, et réussite, et qui fournit généralement des résultats meilleurs que les autre classifieurs d'ensemble comme RF et Rot-For. Elle combine la puissance d'apprentissage des méthodes d'ensemble avec la force de l'analyse de corrélation canonique déjà cités [21,22] et qui a prouvé sa capacité de discrimination pour déterminer les facteurs de risque importants de la récidive du cancer du sein.

Conclusion

Face au très grand nombre de méthodes d'apprentissage statistique présentées dans la littérature, diverses approches sont proposées pour arriver à une classification plus performante parmi elles les méthodes d'ensemble.

Les méthodes d'ensemble consistent à construire plusieurs classifieurs d'une façon différente pour résoudre le problème initial, ou l'objectif visé par ces méthodes est que le prédicteur final soit meilleur que chacun des prédicteurs individuels. On peut obtenir ces classifieurs en introduisant des modifications sur les bases d'apprentissage, l'espace de caractéristiques, les structures des classifieurs, etc.

Actuellement, la forêt de corrélation canonique (CCF) à gagner beaucoup en notoriété dans différents domaines pour résoudre des problèmes de classification et de régression. Son principe basé sur les arbre oblique l'analyse de corrélation canonique garanti la précision de chaque classifieur pour réduire individuellement l'erreur de l'ensemble et la diversité dans l'ensemble pour minimiser la corrélation entre les classifieurs.

Dans le chapitre suivant nous abordons l'application de la forêt de corrélation canonique pour prédire la récidive du cancer du sein.

Chapitre 3

Résultats d'étude : Application de la Forêt de Corrélation Canonique pour la prédiction de la récidive du cancer du sein

Introduction

Nous avons présenté dans le chapitre précédent un aperçu théorique sur le principe des forêts de corrélation canoniques et les déférentes études et travaux qui reposent sur cette méthode.

Dans ce chapitre, après une présentation de la banque de données élaborée dans ce mémoire, nous en viendrons par la suite à l'implémentation de notre approche ainsi que les expérimentations réalisées tous en discutant les résultats obtenus.

1 Présentation de la banque de données « récidive du cancer du sein »

Dans notre projet nous avons utilisé l'ensemble des données créé par Matjaz Zwitter & Milan Soklic sur la récidive du cancer du sein, il provient du référentiel d'apprentissage automatique UCI disponible en ligne [41]. Il est fourni par le centre médical universitaire de l'Institut d'oncologie de Ljubljana (Yougoslavie).

Cet ensemble de données comprend 286 instances répartis en deux classes (récidive du cancer du sein et en non-récidive). Les instances sont décrites par 9 attributs, dont certains sont linéaires et d'autres nominaux. Le tableau Table 3.1 représente les termes standard utilisés dans le référentiel de la banque de données utilisée, la figure 3.1 indique la répartition statistique de chaque paramètre.

PARAMÈTRE	DÉFINITION	VALEURS
Classe	Classe de sortie en fonction de la réapparition des symptômes du cancer du sein chez les patientes après le traitement	no-recurrence-events, recurrence-events
Age	Age de la patiente au moment où la tumeur primaire a été détectée.	[10-99]
menopause	L'état de ménopause de la patiente au moment du diagnostic (pré ou post-ménopausée)	lt40, ge40, pre-meno.
tumeur-size	Décrit la taille de la grosseur qui se forme. La taille de la tumeur est mesurée en millimètre (mm)	[0-59]
inv-node	Indique le nombre de ganglions axillaires qui portent des symptômes de cancer du sein lorsque l'examen histologique est effectué	[0-39]
node-caps	Désigne si la tumeur s'est diffusée dans la capsule ganglionnaire ou non.	yes, no.
deg-malig	Le grade histologique (intervalle 1-3) de la tumeur. Les tumeurs de grade 1 sont principalement constituées de cellules néoplasiques, Les tumeurs de grade 3 sont principalement constituées de cellules qui sont très anormales.	1,2,3.
breast	Le cancer du sein peut évidemment se produire dans l'un ou l'autre des deux seins.	left, right.
Breast-quad	Désigne la région du le sein infecté, elle peut être divisé en quatre quadrants, en utilisant le mamelon comme point central.	left-up, left-low, right-up, right-low, central.
irradiat	Indique si une radiothérapie a été appliquée pour détruire les cellules cancéreuses.	yes, no.

TABLE 3.1 – Description des variables de la base récidive du cancer du sein.

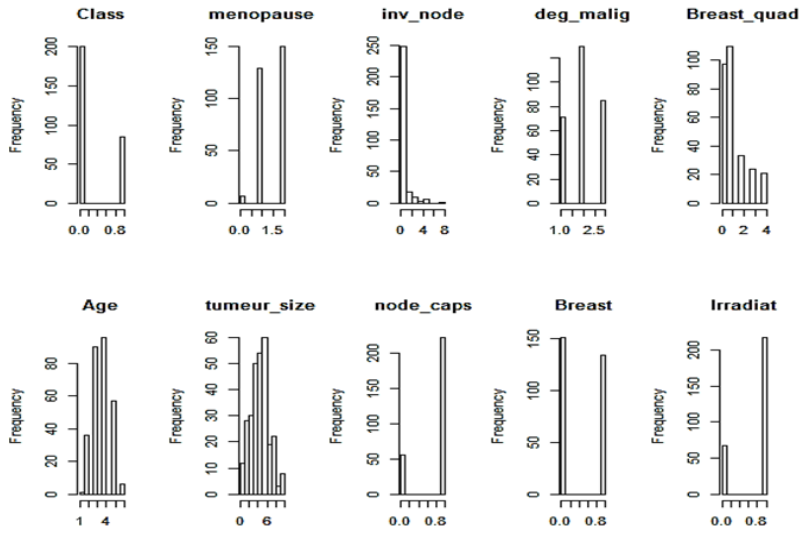

FIGURE 3.1 – Histogrammes des variables de l'ensemble de données.

2 Protocole d'expérimentations

Les méthodes d'ensemble telles que les forêts aléatoires (RF) se base sur la construction d'une collection de prédicteurs et l'agrégation de leurs prédictions, elles sont couramment utilisées pour la classification et la régression par rapport aux techniques classiques en raison de leurs extensibilités, simplicités (en termes de formation et de réglage) et performances. Autrement dit, la précision des arbres individuels et la diversité de leurs prédictions permettent une meilleure tendance pour la prédiction des étiquettes pour les nouvelles données.

L'idée principale de l'algorithme de la forêt de corrélation canonique est de construire un certain nombre d'arbres de corrélation canoniques (CCT). Chaque CCT est construit en appliquant l'analyse de corrélation canonique (ACC) pour trouver les projections de caractéristiques fournissant la corrélation maximale entre les caractéristiques et les étiquettes de classe, puis la meilleure répartition dans cet espace projeté est sélectionnée. Par la suite, pour faire une prévision finale pour des nouveaux échantillons, un vote majoritaire combinant les prévisions des différents CCT est appliqué.

Dans notre travail, l'objectif principal est la prédiction de la récurrence du cancer du sein en effectuant une classification supervisée. De ce fait, nous avons partitionné notre base de données en deux parties, suivant le protocole expérimental de L. Breiman [25] et Rainforth & Wood [27].

Un ensemble d'apprentissage (2/3 de l'ensemble de données) : sont utilisés pour la formation des arbres, et un ensemble de test (1/3 de l'ensemble de données) est utilisé pour une validation croisée afin d'estimer la performance du modèle. L. Breiman [25] a défini deux propriétés pour l'erreur de généralisation d'une forêt qui dépend de la force des arbres individuels permettant de mesu-

rer la fiabilité de la forêt et de la corrélation entre eux qui représente le degré de dépendance des arbres de la forêt.

2.1 Choix des paramètres d'algorithmes

Il existe deux principaux paramètres dans la méthode des forêts aléatoire (RF) :

1. Le paramètre le plus important est le nombre de variables choisies aléatoirement à chacun des nœuds des arbres $mtry$ (Il peut varier de 1 à p (le nombre de variables dans la base de données) et possède une valeur par défaut : $\sqrt{p}$ en classification, $p3$ en régression).
2. Le deuxième paramètre est le nombre d'arbres dans la forêt. Il est nommé $ntree$. En effet, Le choix le plus judicieux étant après plusieurs expérimentations.

Dans le même concept, sachant que CCF est une méthode non paramétrique, de ce fait les mêmes paramètres de RF sont appliqués aux forêts de corrélation canonique pour la prédiction du cancer du sein.

2.2 Critère d'évaluation

Le critère d'évaluation est un facteur clé pour évaluer la performance de classification et guider la modélisation de classifieur. Pour une comparaison synthétique des différentes méthodes de classification, nous avons calculé : Taux de classification : c'est le pourcentage des exemples correctement classés, donné par :

$$TC = [100 * (VP + VN)/(VN + VP + FN + FP)]$$

Dans notre cas :

- VP ou vrai positif : récurrent classé récurrent
- FP ou faux positif : non-récurrent classé récurrent
- VN ou vrai négatif : non-récurrent classée non-récurrent.
- FN ou faux négatif : récurrent classée non-récurrent.

Dans cette étude, nous voulons vérifier la pertinence des forêts de corrélation canonique CCF en comparaison aux Foret Aléatoire RF et Rotationnelle Rot-For, qui a démontré sa supériorité dans les travaux d'état de l'art en termes de performance et robustesse ainsi que la rapidité du temps de calcul.

Concernant, le choix du nombre d'arbres, nous proposons de faire des tests avec un nombre d'arbres variant de 5 à 100 en utilisant 5 validations croisées pour l'ensemble de données.

Les langages de programmation utilisé sont : MATLAB R2017a pour les différents algorithmes RF, Rot-For et ccfs-master toolbox [1], SPSS 25 pour applications des analyses des données.

1. https ://github.com/twgr/ccfs

3 Résultats et Discussion

La synthèse des résultats obtenus avec le protocole expérimentation détaillé est présentée dans le tableau Table 3.2 et la figure 3.2, avec des concours comparatifs organisés avec les forets aléatoires (RF) et la forêt de rotation (Rot-For) afin d'étudier les performances prédictives de la forêt de corrélation canonique.

Les résultats montrent clairement que les CCF surpassent significativement les deux autres méthodes avec une valeur maximale de reconnaissance égale 72.29% pour un nombre d'itérations de 50 et 20.

Nbre d'arbres	Forêt Aléatoire (RF)	Forêt Rotationnelle (Rot-For)	Forêt de Corrélation Canonique (CCF)
5	0.6813	0.6708	0.7042
10	0.6958	0.6917	0.7042
15	0.6875	0.6583	0.7063
20	0.7169	0.7063	0.7229
25	0.6854	0.6813	0.6854
30	0.7104	0.7125	0.7146
40	0.6979	0.6686	0.7
50	0.7063	0.6813	0.7229
60	0.7125	0.6958	**0.7271**
80	0.7118	0.691	0.7222
100	0.6664	0.6396	0.6854

Table 3.2 – Performances des différents classifieurs.

Nous postulons que les CCF sont meilleures que RF et Rot-For ce qui est de l'analyse de corrélation canonique, elles représentent une nouvelle référence de performance pour les ensembles d'arbres de décision, en tenant compte de ces points, ils ont fourni des preuves empiriques substantielles suggérant qu'ils surpassent un certain nombre d'approches importantes telles que RF et Rot-For.

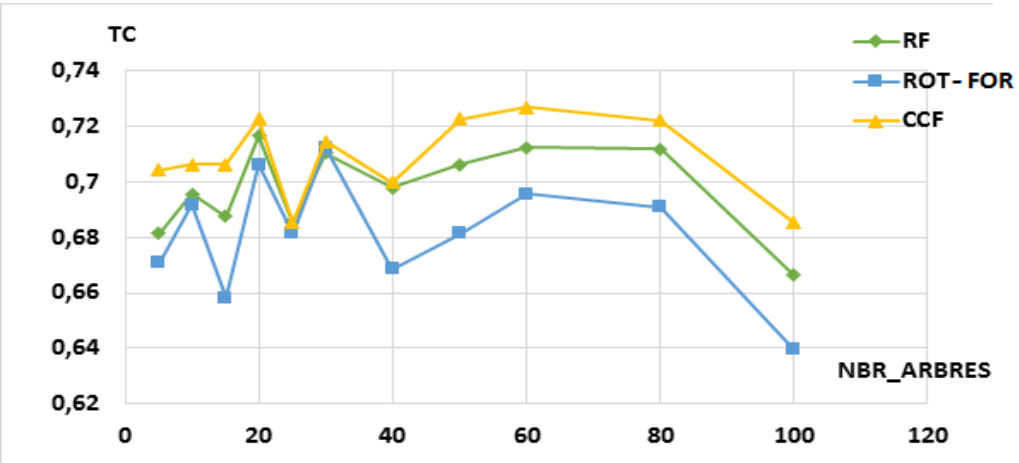

Figure 3.2 – Taux de classification en fonction du nombre d'arbres.

Pour une étude comparative des résultats obtenus dans le tableau 3.1, certaines points doivent être clarifiés et certains d'autres sont constatés pour les différents algorithmes.

3.1 L'algorithme de Random Forest

L'algorithme RF est la base des autre algorithmes sans aucune amélioration, les résultats que nous avons obtenues avec $mtry$ fixé à 3 et un élagage des arbres, pour éviter le sur apprentissage, dans le cas ou $ntree$ atteins les 25 arbres, nous remarquons qu'il donne la même valeur de taux de classification avec CCF.

3.2 L'algorithme de Rotation Forest

D'après ses résultats, nous constatons une petite amélioration par rapport aux RF avec une valeur maximal de reconnaissance égale 71.25% pour un nombre des itérations égal à 30 dont l'explication heuristique de cette dernière revient à la rotation des arbres, ainsi l'application de l'analyses des composant principales (ACP), d'autre part, leur robustesse semble similaire aux corrélations globales. L'analyse en Composantes Principales (ACP) est une méthode d'analyse de données. Elle cherche à synthétiser l'information contenue dans un tableau croisant des individus et des variables quantitatives. Produire un résumée d'information, au sens de l'ACP c'est établir une similarité entre les individus, chercher des groupes d'individus homogènes, mettre en évidence une typologie d'individus.

Nous avons appliqué ACP sur notre base de données sur SPSS et les résultats sont illustrés dans les tableaux Table 3.3 et 3.4.

	Extraction sommes des carrés des facteurs retenus		
Composantes	Total	% de la variance	% cumulés
1	2.228	24.752	24.752
2	1.676	18.628	43.380
3	1.174	13.043	56.423

Table 3.3 – Les composantes principales.

variables	Extraction
Age	.819
menopause	.829
tumeur_size	.428
inv_node	.664
node_caps	.663
deg_malig	.384
Breast	.355
Breast_quad	.578
irradiat	.359

Table 3.4 – La qualité de représentation des variables.

La tableau Table 3.4 présente la qualité de représentation et en nous basons sur ses valeurs nous remarquons clairement que : menopause, Age, inv_node, node_caps, Breast_quad et tumeur-size respectivement sont des variables importantes dans cette étude ce qui explique l'apport de ces derniers dans la projection des variables pour l'amélioration des performances.

3.3 L'algorithme de Canonical Correlation Forest

Il est intéressant de noter que la performance des CCF a sensiblement améliorée la précision prédictive du cancer du sein par rapport aux autres méthodes, en tenant compte de ces résultats, le facteur clé de ces améliorations repose sur les arbre oblique et l'application de l'analyse de corrélation canonique (ACC).

L'analyse de corrélation canonique (ACC) est une méthode d'évaluation de corrélation linéaire entre deux ensembles de variables, c'est-à-dire elle permet d'examiner la relation entre eux qui peuvent identifier des variables importantes dans un ensemble des prédicteurs et un ensemble de résultats multiples. Dans notre cas, l'ACC est utilisée pour la division d'hyperplan définies par une combinaison des caractéristiques, basés sur les coefficients de corrélation canoniques locaux calculés pendant l'apprentissage.

Après une application de l'ACC, les résultats sont illustrés dans les tableaux Table 3.5, 3.6 et 3.7. Une façon d'interpréter les solutions canoniques consiste à examiner les corrélations entre les variables canoniques et les variables dans chaque ensemble. Ces corrélations sont appelées coefficients de structure (loading).

Le critère de choix des variables importantes dans chaque variable canonique sont les coefficients de structure. En règle générale, pour les valeurs significatives, une valeur absolue égale à ou supérieure à 0,3 est souvent utilisée. D'autre part, les signes les coefficients de structure aident à identifier le caractère de la relation entre les variables du prédicteur et des ensembles de résultats. Si les deux ont le même signe, alors ils changent dans le même sens ; et vice versa [22].

L'utilisation des coefficients des structure de l'ACC facilite la détection des prédicteurs, en particulier lorsqu'il existe de nombreuses variables dans l'ensemble de données et il existe des corrélations élevées entre ces variables.

	CORRELATION	EIGENVALUE	WILKS STATISTIC	F	NUM D.F	DENOM D.F	SIG
1	.425	.219	.806	3.801	16.000	534.000	.000
2	.131	.018	.983	.670	7.000	268.000	.697

TABLE 3.5 – Exemple de cas des variables canoniques.

Les variables dans les deux ensembles sont classées selon les valeurs absolues des charges, ces derniers montrent leur importance au sein de chaque variable canonique.

VARIABLES	1	2
Age	.207	.333
menopause	.209	-.466
tumeur_size	-.448	-.413
inv_node	-.698	.059
node_caps	-.681	-.045
deg_malig	-.777	.294
Breast_quad	.036	.685
irradiat	.534	.231

TABLE 3.6 – Exemple de cas des coefficients de structure des variables du premier ensemble.

VARIABLES	1	2
BREAST	.022	-1.000
CLASSE	-1.000	.019

TABLE 3.7 – Exemple de cas des coefficients de structure des variables de deuxième ensemble

Selon les résultats illustrés dans les tableaux Table 3.6 et 3.7, nous remarquons clairement que : le grade histologique de la tumeur (deg_malig), le nombre de ganglions axillaires qui portent des symptômes de cancer du sein (inv_node), indicateur de diffusion de la tumeur dans la capsule ganglionnaire (node_caps), indice d'application du radiothérapie (irradiat), la taille de la tumeur (tumeur-size) et le signe d'apparition du cancer au niveau d'un des deux sein (breast) sont des variables importantes pour la récidive de cancer du sein. Nous supposons que les variables : Age, menopause et Breast-quad ne sont pas importants en tant que prédicteurs de la récurrence de la maladie.

Pour expliquer la supériorité des CCF aux forêts de rotation, nous postulons les points suivants :

- L'incorporation des corrélations localisées puisque les plans d'hyperplan sont calculés à chaque nœud séparément, donc les CCF sont plus efficaces en incorporant des corrélations locales.
- En outre, les CCF sont dépendantes de la classe, cela est dû au fait que l'étape de rotation des forêts de rotation n'intègre aucune information sur les classes, sauf dans l'élimination aléatoire des classes.
- De plus, les arbres individuels d'une forêt de rotation sont orthogonaux et ne peuvent donc pas incorporer de variation spatiale dans la corrélation. La nature auto-suffisante de l'algorithme de croissance pour les arbres de corrélation canonique signifie que les corrélations locales d'une partition peuvent être incorporées aussi naturellement que les corrélations globales.

Aussi en terme du temps, CCF comme toute méthodes d'ensemble, elles demandent un temps d'exécution un peu long avec une relation proportionnelle

avec le nombre d'itération, mais cela est seulement dans la phase d'apprentissage, en phase de test on remarque qu'il nous propose un gain de temps très important.

4 Comparaison générale

Pour le même objectif lequel est la prédiction de récidive de cancer du sein, nombreux travaux sont mis en place tels que les deux cités dans le chapitre 1 [6, 16] qui se basent sur la même base de données :

N. Alva [6] a évalué plusieurs algorithmes de classification d'apprentissage automatique après un pré-traitement des données, et ils ont trouvé que l'algorithme de régression logistique et Naïve Bayes avaient la plus grande précision à 73%.

En outre, Goyal et al. [16] 2020 ont testé divers classifieurs de classification et les résultats ont montré que réseaux de neurones surpassent les autres avec un taux de reconnaissance égal à 85.18%.

En revanche, dans notre travail nous sommes arrivés à un taux maximal égal à 72.71% dont les arguments que nous jugeons appréciable dans les méthodes d'ensemble sont la précision de chaque arbre individuel et la diversité de prédiction qui offrent plus de robustesse à l'approche.

5 Conclusion

Dans ce chapitre, nous avons abordé l'application des forêts de corrélation canoniques (CCF), une nouvelle méthode d'ensemble d'arbres de décision performante qui fournit généralement des résultats meilleurs que les autre classifieurs d'ensemble, pour classification, la régression et la prédiction de sorties multiples.

La clé de réussite et de succès apportées aux CCF par rapport aux précédentes méthodes d'ensemble repose sur l'utilisation de l'analyse de corrélation canonique, elle est considérée comme une nouvelle référence de performance en tenant compte qu'elle donne des meilleurs résultats en comparant avec d'autres techniques de classification.

Conclusion

Les algorithmes d'apprentissage automatique peuvent être utilisés pour compléter le modèle de soins existant pour un diagnostic médical. Dans la littérature, un certain nombre d'approches différentes ont été conçues afin de chercher une précision ainsi qu'une efficacité parfaite, à titre d'exemple les méthodes d'ensemble.

Durant ces dernières années, les méthodes d'ensemble sont devenues l'un des axes de recherches les plus populaires car ils permettent d'ajouter le comportement global au comportement individuel ce qui améliore de manière significative les performances obtenues en classification par rapport aux méthodes classiques c'est à dire l'utilisation d'un classifieur unique.

Le travail effectué dans le cadre de ce projet de fin d'études a concerné l'évaluation de l'une des méthodes d'ensemble d'arbre nommé *"forêt de corrélation canonique"* comme alternative aux forêts aléatoires bien connue pour leur pouvoir prédictif, les propulsant comme candidat idéal à la prédiction de la récidive du cancer du sein.

Le choix de cette méthode a été fait après une étude bibliographique en tenant compte qu'elle représente une technique de génération d'ensemble d'arbre indépendant la plus performante, et qui fournit généralement des résultats meilleurs que les autres classifieurs d'ensemble. En plus facile à mettre en œuvre et fonctionne d'une manière très efficace. L'un des avantages les plus significatif de la méthode proposée est qu'elle ne nécessite pas de réglage de paramètres et elle se base principalement sur l'analyse de corrélation canonique pour la construction de l'ensemble classifieurs qui porte une amélioration simultanée sur la précision individuelle et la diversité au sein de l'ensemble.

Nous nous sommes intéressés dans ce travail à étudier la performance de cette méthode en comparaison avec les forêts aléatoires et les forêts rotationnelles sur la base de données « récidive du cancer du sein » pour l'évaluation, basé sur la validation croisée, les résultats obtenus ont montré que :

- La méthode proposée fourni les meilleurs résultats en les comparant avec les deux citées, plus performante, elle surpasse les autres pour la récidive du cancer du sein ce qui valide notre contribution de départ.

- Les variables le grade histologique de la tumeur (deg_malig), le nombre de

ganglions axillaires qui portent des symptômes de cancer du sein (inv_node), indicateur de diffusion de la tumeur dans la capsule ganglionnaire (node_caps), indice d'application du radiothérapie (irradiat), la taille de la tumeur (tumeur-size) et le signe d'apparition du cancer au niveau d'un des deux sein (breast) sont les variables intervenantes pour la récidive de cancer du sein..

Dans la perspective de mettre en place un certain nombre d'études pour mieux développer cette approche dans le futur le plus proche, tout en proposant d'utiliser le SMOTE (Synthetic Minority Over-sampling Technique) sachant que la base de données est déséquilibrée, avec un nombre significativement plus élevé de cas sans récidive par rapport aux cas de récidive.

Bibliographie

[1] ALGÉRIE PRESSE SERVICE, "Cancer du sein," http ://www.aps.dz/sante-science-technologie/96195-cancer-du-sein-le-depistage-precoce-de-nouveau-preconise, Consulté le 07/08/2020.

[2] Le Figaro, "Cancer du sein," https ://sante.lefigaro.fr/sante/maladie/cancer-sein/quest-ce-que-cest, Consulté le 20/02/2020.

[3] Association of Breast Surgery, "Recurrent and metastatic breast cancer data collection project," Tech. Rep., Breast Cancer Care, 2012.

[4] Institut National Du Cancer, "Facteurs-de-risque-de-recidive," https ://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Facteurs-de-risque-de-recidive, Consulté le 13/02/2020.

[5] Erika J Schneble, Lindsey J Graham, Matthew P Shupe, Frederick L Flynt, Kevin P Banks, Aaron D Kirkpatrick, Aviram Nissan, Leonard Henry, Alexander Stojadinovic, Nathan M Shumway, et al., "Current approaches and challenges in early detection of breast cancer recurrence," *Journal of Cancer*, vol. 5, no. 4, pp. 281, 2014.

[6] Nandita Alva, "Using machine learning techniques to predict the recurrence of breast cancer," https ://github.com/NanditaA/ML-Supervised-Classification/blob/master/Breast2018.

[7] Pedro Henriques Abreu, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, and Daniel Castro Silva, "Predicting breast cancer recurrence using machine learning techniques : A systematic review," *ACM Comput. Surv.*, vol. 49, no. 3, Oct. 2016.

[8] Ravi Aavula and Bhramaramba Ravi, "Xbpf : An extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability," in *International Journal of Engineering and Advanced Technology (IJEAT)*, 06 2019, vol. Volume-8 of *ISSN : 2249-8958,*.

[9] Desta Mulatu and Rupali R. Gangarde, "Survey of data mining techniques for prediction of breast cancer recurrence," *International Journal of Computer Science and Information Technologies (IJCSIT),*, vol. 8, pp. 3–11, 2017.

[10] Linda Aagaard Rasmussen, Henry Jensen, Line Flytkjaer Virgilsen, Lise Bech Jellesmark Thorsen, Birgitte Vrou Offersen, and Peter Vedsted, "A validated algorithm for register-based identification of patients with recurrence of breast cancer-based on danish breast cancer group (dbcg) data," *Cancer Epidemiology*, vol. 59, pp. 129 – 134, 2019.

[11] Mohammad R. Mohebian, Hamid R. Marateb, Marjan Mansourian, Miguel Angel Mananas, and Fariborz Mokarian, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (hpbcr) using optimized ensemble learning," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 75 – 85, 2017.

[12] Mariko Asaoka, Kazutaka Narui, Nobuyasu Suganuma, Takashi Chishima, Akimitsu Yamada, Sadatoshi Sugae, Saori Kawai, Natsuki Uenaka, Saeko Teraoka, Kana Miyahara, et al., "Clinical and pathological predictors of recurrence in breast cancer patients achieving pathological complete response to neoadjuvant chemotherapy," *European Journal of Surgical Oncology*, vol. 45, no. 12, pp. 2289–2294, 2019.

[13] Chi-Chang Chang and Ssu-Han Chen, "Developing a novel machine learning-based classification scheme for predicting spcs in breast cancer survivors," *Frontiers in Genetics*, vol. 10, pp. 848, 2019.

[14] L Gh Ahmad, AT Eshlaghy, A Poorebrahimi, M Ebrahimi, AR Razavi, et al., "Using three machine learning techniques for predicting breast cancer recurrence," *J Health Med Inform*, vol. 4, no. 124, pp. 3, 2013.

[15] Esther Paredes-Aracil, Antonio Palazón-Bru, David Manuel Folgado-de la Rosa, José Ramón Ots-Gutiérrez, Cristina Llorca-Ferrándiz, Sonia Alonso-Hernández, José Vicente Coloma-Lidón, and Vicente Francisco Gil-Guillén, "A scoring system to predict recurrence in breast cancer patients," *Surgical oncology*, vol. 27, no. 4, pp. 681–687, 2018.

[16] Kashish Goyal, Preeti Aggarwal, and Mukesh Kumar, "Prediction of breast cancer recurrence : A machine learning approach," in *Computational Intelligence in Data Mining*, Himansu Sekhar Behera, Janmenjoy Nayak, Bighnaraj Naik, and Danilo Pelusi, Eds., Singapore, 2020, pp. 101–113, Springer Singapore.

[17] Uma Ojha and Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques," in *7th International Conference on Cloud Computing, Data Science Engineering - Confluence*, Jan 2017, pp. 527–530.

[18] Ahmed Iqbal Pritom, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in *19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2016, pp. 310–314.

[19] Robin Stuart-Harris, Jane E Dahlstrom, Ruta Gupta, Yanping Zhang, Paul Craft, and Bruce Shadbolt, "Recurrence in early breast cancer : Analysis of data from 3,765 australian women treated between 1997 and 2015," *The Breast*, vol. 44, pp. 153–159, 2019.

[20] Yu-Hui Zhou, Yang Liu, Wei Zhang, Chao Liu, Jian-Jun He, and Xiao-Jiang Tang, "Associations between clinical-pathological parameters and biomarkers, her-2, tyms, rrmi, and 21-gene recurrence score in breast cancer," *Pathology-Research and Practice*, vol. 215, no. 11, pp. 152–644, 2019.

[21] Farahnaz Sadoughi, Hadi Lotfnezhad Afshar, Asiie Olfatbakhsh, and Neda Mehrdad, "Application of canonical correlation analysis for detecting risk

factors leading to recurrence of breast cancer," *Iranian Red Crescent Medical Journal*, vol. 18, no. 3, 2016.

[22] Amir R Razavi, Hans Gill, Olle Stål, Marie Sundquist, Sten Thorstenson, Hans Åhlfeldt, Nosrat Shahsavar, South-East Swedish Breast Cancer Study Group, et al., "Exploring cancer register data to find risk factors for recurrence of breast cancer–application of canonical correlation analysis," *BMC medical informatics and decision making*, vol. 5, no. 1, pp. 29, 2005.

[23] Leo Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[24] Yoav Freund and Robert E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, San Francisco, CA, USA, 1996, ICML96, p. 148156, Morgan Kaufmann Publishers Inc.

[25] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] Juan Rodriguez, Ludmila Kuncheva, and Carlos Alonso, "Rotation forest : A new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1619–30, 11 2006.

[27] Tom Rainforth and Frank Wood, "Canonical correlation forests," *arXiv preprint arXiv :1507.05444*, 2015.

[28] Camille Duby and Stéphane Robin, "Analyse en composantes principales," *Institut National Agronomique, Paris-Grignon*, vol. 80, 2006.

[29] Anthony J. Bagnall, Aaron Bostrom, Gavin C. Cawley, Michael Flynn, James Large, and Jason Lines, "Is rotation forest the best classifier for problems with continuous features ?," *CoRR*, vol. abs/1809.06705, 2018.

[30] Tin Kam Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, USA, 1995, ICDAR 95, p. 278, IEEE Computer Society.

[31] Harold Hotelling, *Relations Between Two Sets of Variates*, pp. 162–190, Springer New York, New York, NY, 1992.

[32] Alston S. Householder, "Unitary triangularization of a nonsymmetric matrix," *J. ACM*, vol. 5, no. 4, pp. 339342, Oct. 1958.

[33] Gene Howard Golub and Christian H Reinsch, "Singular value decomposition and least squares solutions," *Numer. Math.*, vol. 14, no. 5, pp. 403420, Apr. 1970.

[34] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim, "Do we need hundreds of classifiers to solve real world classification problems ?," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 31333181, Jan. 2014.

[35] Ha Thang, Merilyn Manley-Harris, Tien Dat Pham, and Ian Hawes, "A comparative assessment of ensemble-based machine learning and maximum likelihood methods for mapping seagrass using sentinel-2 imagery in tauranga harbor, new zealand," *Remote Sensing*, vol. 12, pp. 355, 01 2020.

[36] Ismail Colkesen and Taskin Kavzoglu, "Ensemble-based canonical correlation forest (ccf) for land use and land cover classification using sentinel-2 and landsat oli imagery," *Remote Sensing Letters*, vol. 8, no. 11, pp. 1082–1091, 2017.

[37] Junshi Xia, Naoto Yokoya, and Akira Iwasaki, "Hyperspectral image classification with canonical correlation forests," *IEEE Transactions on Geoscience and Remote Sensing*, 09 2016.

[38] Xuanhui Wang, Tailian Liu, Xilai Zheng, Hui Peng, Jia Xin, and Bo Zhang, "Short-term prediction of groundwater level using improved random forest regression with a combination of random features," *Applied Water Science*, vol. 8, 09 2018.

[39] Daniel T Ocansey, Marvin Aidoo, Marwan Bikdash, Hamid D Ismail, Clarence White, Robert H Newman, and B KC Dukka, "Performance of canonical correlation forest in phosphorylation site predictions," in *SoutheastCon 2018*. IEEE, 2018, pp. 1–7.

[40] Emrehan Kutlug Sahin, Ismail Colkesen, and Taskin Kavzoglu, "A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping," *Geocarto International*, vol. 35, no. 4, pp. 341–363, 2020.

[41] Matjaz Zwitter and Milan Soklic, "Breast cancer data set," archive.ics.uci.edu, Access–February 2020, `https://archive.ics.uci.edu/ml/datasets/breast+cancer`.

Résume : La récidive est la réapparition d'un cancer à partir de cellules cancéreuses non détruites par le traitement initial. Ce phénomène ne ressurgit pas à une période bien précise, la possibilité de récurrence du cancer est alors une situation éprouvante à vivre, pour cela aujourd'hui, nombreuses sont les méthodes d'apprentissage automatique qui ont été appliquées pour améliorer les performances et augmenter l'efficacité des systèmes de prédiction de la récurrence du cancer. Durant cette dernière décennie, les chercheurs se sont plus intéressés aux méthodes d'ensemble tenant compte de leur grande précision, grâce à leur robustesse et capacité à préserver l'information de variabilité des données. Parmi elles : la forêt aléatoire, forêt rotationnelle et la forêt de corrélation canonique. Dans notre projet de fin d'études, nous proposons d'utiliser les forêts de corrélation canoniques (CCF), une nouvelle méthode d'ensemble d'arbres de décision pour la prédiction de la récidive du cancer du sein. Cette méthode combine la puissance d'apprentissage des méthodes d'ensemble avec la force de discrimination de l'analyse canonique tout en augmentant la précision de chaque arbre et la diversité des arbres dans la forêt. Les résultats expérimentaux appliqués sur la banque de données médicales récidive du cancer du sein montrent une amélioration de performance dans la tâche de la prédiction en comparaison avec la forêt aléatoire et la forêt rotationnelle en termes de taux de classification comme critère d'évaluation des performances.

Mots clés : la récurrence du cancer du sein, prédiction, méthodes d'ensembles, forêt aléatoire, forêt rotationnelle, forêts de corrélation canonique, analyse de corrélation canonique, bases de données UCI.

Abstract : Recurrence is the re-emergence of cancer from cancer cells not destroyed by the initial treatment. This phenomenon does not reappear at a specific period of time, the possibility of cancer recurrence is then a trying situation to live through, for this reason today, many automatic learning methods have been applied to improve the performance and increase the efficiency of cancer recurrence prediction systems. During the last decade, researchers have become more interested in ensemble methods considering their high precision, because of their robustness and ability to preserve the information of data variability. These include Random Forest, Rotational Forest and Canonical Correlation Forest. In this Master Thesis, we propose to use Canonical Correlation Forests (CCF), a new decision tree ensemble method for predicting breast cancer recurrence. This method combines the learning power of ensemble methods with the discriminating power of canonical analysis while increasing the accuracy of individual trees and the diversity of trees in the forest. Experimental results applied to the breast cancer recurrence dataset show an improvement in performance in the prediction task compared to Random Forest and Rotational Forest in terms of classification rate as a performance evaluation criterion.

Keywords : breast cancer recurrence, prediction, ensemble methods, random forest, rotational forest, canonical correlation forests, canonical correlation analysis, databases.

ملخص: تكرار الإصابة بسرطان الثدي هو عودة السرطان بسبب ظهور الخلايا السرطانية التي لم يتم تدميرها بالعلاج الاول من جديد. هذه الظاهرة لا تظهر مرة أخرى في وقت محدد، وبالتالي فإن احتمالية تكرار الإصابة بالسرطان هي حالة صعبة للعيش، لهذا في يومنا هذا، تم تطبيق العديد من طرق التعلم الالي لتحسين الأداء وزيادة كفاءة أنظمة دعم التشخيص الطبي. خلال العقد الماضي، أصبح الباحثون أكثر اهتمامًا بأساليب المجموعات مع الأخذ بعين الاعتبار دقة نتائجهم نظرا لقوتهم وقدرتهم على الحفاظ على المعلومات وتنوع البيانات. من بينها: الغابة العشوائية، الغابة الدورانية، وغابة الارتباط القانوني .في مشروع التخرج الخاص بنا، نقترح استخدام غابات الارتباط القانوني، وهي طريقة جديدة تابعة لأساليب مجموعات الاشجار التي تم تعيينها للتنبؤ بتكرار الإصابة بسرطان الثدي. تجمع هذه الطريقة بين القوة التعليمية لطرق التجميع والقوة التمييزية للتحليل القانوني مع زيادة دقة الأشجار الفردية وتنوع الأشجار في الغابة. أظهرت النتائج التجريبية المطبقة على بنك البيانات الطبية (تكرار الإصابة بسرطان الثدي) أن غابات الارتباط القانوني أحسن أداءا لعملية التنبؤ مقارنة مع الغابة العشوائية والغابة الدورانية من حيث معدل التصنيف كمعيار للتقييم.

كلمات البحث: تكرار الإصابة بسرطان الثدي، التنبؤ، أساليب المجموعات ، الغابة العشوائية ، الغابة الدورانية ، غابة الارتباط القانوني، التحليل القانوني ، قواعد البيانات.