Ministère de l'enseignement Supérieur et de la recherche Scientifique
وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculté des Technologie

Département d'informatique

جامعة باجي مختار – عنابـــــــــة
كلية التكنولوجيا
قســـــــم الإعلام الآلي

Thèse

Présentée pour obtenir le diplôme de Doctorat en LMD

Doctorat

Spécialité : Intelligence artificielle et traitement d'image

Par :
Ilhem Tarchoune

Thème :

Utilisation des forêts aléatoires pour améliorer les performances d'un système de Raisonnement à Partir de Cas médical

N°	Nom et prénom	Grade	Etablissement	Qualité
01	Toufik Sari	Prof	Université Badji Mokhtar –Annaba	Président
02	Akila Djebbar	MCA	Université Badji Mokhtar –Annaba	Rapporteur
03	Nabiha Azizi	Prof	Université Badji Mokhtar –Annaba	Examinateur
04	Fayez Mazouzi	MCA	Université Mohamed-Chérif Messaadia - Souk Ahras	Examinateur
05	Abedlmadjid Benmachiche	MCA	Université Chadli Bendjedid-Tarf	Examinateur
06	Hayet Farida Merouani	Prof	Université Badji Mokhtar –Annaba	Invité

Remerciements

Je souhaite d'abord exprimer ma gratitude envers Dieu tout puissant pour m'avoir accordé le courage, la force et la patience nécessaires pour mener à bien ce travail.

Je tiens également à exprimer ma profonde reconnaissance à ma directrice de thèse, le Docteur **DJEBBAR Akila**, pour son encadrement, ses conseils inestimables, sa bonne humeur, ainsi que pour son aide précieuse dans la révision de ma thèse. Elle a toujours été attentive et disponible, malgré ses nombreuses responsabilités.

Je remercie sincèrement le Professeur **MEROUANI Hayette Farida** pour avoir dirigé ma thèse avec dévouement et patience.

Je tiens également à exprimer toute ma reconnaissance au Professeur **SARI Toufik** pour avoir accepté de présider le jury de cette thèse.

Enfin, mes remerciements vont également aux Professeurs **AZIZI Nabiha** et **MAZOUZI Fayez** pour l'honneur qu'ils m'ont fait en acceptant de faire partie de mon jury de thèse.

Dédicace

Mes grands remerciements sont pour notre Dieu qui m'a aidé et m'a donné le pouvoir, la patience et la volonté d'avoir réalisé ce travail.

Je dédie mon travail à mes chers parents, que j'aime tant et qui m'ont toujours encouragé avec une inéluctable patience pendant mes langues études. Qu'ils trouvent ici le témoignage de ma gratitude, leurs affections, leurs amours, leurs encouragements et leurs sacrifices ; qu'ils n'ont pas cessé de me procurer durant mes études.

A mon mari, Merci pour votre amour, vos encouragements, Que Dieu vous garde.
A mes enfants chéris : Amine, Djoud, Chahd, Que Dieu vous garde.

Ces dédicaces vont également : A mes chères sœurs : Amel et Nour et mes nièces : Maram, Rym et Maria.
A mes beaux-frères : Samir, Ahmed et Nisar.

A toute ma famille Tarchoune et Karkour, mes oncles et mes tantes et leurs familles.

A mes chers amis(es) : Hadjer, Sarra, Oulaya, avec les quels (elles) j'ai partagé mes meilleures années d'études.

A mon directeur de thèse : Akila DJEBBAR

A mes collègues de l'université Badji Mokhtar Annaba. Et à tous ceux qui m'ont apporté d'aide de près ou de loin.

Résumé

Le raisonnement à partir de cas (RàPC) est une approche de résolution de problème basée sur des expériences passées appelées ‘cas’. La qualité de cette approche est directement liée au développement de différentes étapes ainsi la qualité de leurs bases de cas, ce qui rend la modification de ces dernières d’une grande importance. Nous sommes intéressé à la modification de la phase de la remémoration, qui est un composant clé dans le cycle RàPC et l’adaptation du cas remémoré qui est considéré une des étapes les plus difficiles du cycle RàPC.

L’idée principale est de modéliser la phase de remémoration par une forêt aléatoire améliorée. Les forêts aléatoires sont des excellents outils de classification dans le domaine médical. En effet, la forêt aléatoire est considérée comme une mesure de similarité dans la phase de remémoration qui sert à sélectionner le cas le plus similaire. Ensuite, la phase d’adaptation consiste à adapter le cas remémoré selon le problème donné.

Notre objectif consiste à garantir la recherche d’un cas qui soit le plus facile à adapter afin d’améliorer la performance du RàPC. Les performances de l’utilisation des forêts aléatoires (FA) améliorées sont évaluées sur plusieurs bases de données médicales. Les résultats obtenus sont satisfaisants et très encourageants et montrent l’efficacité des forêts aléatoires comme outils de classification et de raisonnement diagnostique.

Mots clé : Raisonnement à Partir de Cas (RàPC), Forêt aléatoire (FA), Sélection de caractéristiques, Remémoration, Réutilisation (Adaptation), Données médicales.

Abstract

Case-Based Reasoning (CBR) is a problem-solving approach based on past experiences called 'cases'. The quality of this approach is directly related to the development of different stages and the quality of their case base, which makes the modification of the latter very important. We are interested in modifying the retrieval phase, which is a key component in the CBR cycle, and adapting the retrieved case, which is considered one of the most difficult stages of the CBR cycle.

The main idea is to model the retrieval phase using an enhanced random forest. Random forests are excellent classification tools in the medical field. Indeed, the random forest is considered a measure of similarity in the retrieval phase, which is used to select the most similar case. Then, the adaptation phase consists of adapting the retrieved case according to the given problem.

Our objective is to ensure the search for a case that is easiest to adapt in order to improve the performance of CBR. The performance of the use of enhanced random forests (RF) is evaluated on several medical databases. The results obtained are satisfactory and very encouraging, showing the effectiveness of random forests as classification and diagnostic reasoning tools.

Keywords: Case-Based Reasoning (CBR), Random Forest (RF), Feature Selection (FS), Retrieval phase, Adaptation phase, Medical data.

الملخص

الاستدلال القائم على الحالات (CBR) هو نهج لحل المشكلات يعتمد على التجارب السابقة المسماة "حالات". يرتبط جودة هذا النهج مباشرة بتطوير مراحل مختلفة وجودة قاعدتهم من الحالات، مما يجعل تعديل الأخيرة أمرًا مهمًا للغاية. نحن مهتمون بتعديل مرحلة الاسترجاع، والتي هي عنصر رئيسي في دورة (CBR) وتكييف الحالة المسترجعة، والتي تعتبر واحدة من أصعب مراحل دورة (CBR) .

الفكرة الرئيسية هي تصميم مرحلة الاسترجاع باستخدام غابة عشوائية محسنة. الغابات العشوائية هي أدوات تصنيف ممتازة في المجال الطبي. في الواقع يُعتبر الغابة العشوائية مقياسًا للتشابه في مرحلة الاسترجاع والتي تستخدم لتحديد الحالة الأكثر تشابهًا, ثم تتكون مرحلة التكييف من تكييف الحالة المسترجعة وفقًا للمشكلة المعطاة.

هدفنا هو ضمان البحث عن حالة يكون تكييفها أسهل لتحسين أداء (CBR).يتم تقييم أداء استخدام الغابات العشوائية المحسنة (RF) على عدة قواعد بيانات طبية, النتائج المتحققة مرضية ومشجعة للغاية، مما يظهر فعالية الغابات العشوائية كأدوات للتصنيف والاستدلال التشخيصي.

الكلمات الرئيسية: الاستدلال القائم على الحالات (CBR)، الغابة العشوائية (RF)، اختيار السمات (FS)، الاسترجاع، التكييف، البيانات الطبية.

Table des matières

Chapitre 3. Techniques de sélection des caractéristiques

Chapitre 4. Contributions

Liste des tableaux

Liste des figures

Liste des Abréviations

IA	Intelligence Artificielle
RàPC	Raisonnement à partir de cas
CBR	Case Based Reasoning
BC	Base de cas
Pb	Problème
KPPV	K plus proche voisine
KNN	K-Nearest Neighbord
RF	Random Forest
OOB	Out of Bag
FS	Feature Selection
AG	Algorithme Génétique
UCI	Université de Californie à Irvine
LMT	Logic Model Tree
TC	Taux de classification
Se	Sensitivité
Sp	Spécificité
TE	Taux d'erreur

Introduction Générale

Contexte et problématique

L'intelligence artificielle (IA) vise à reproduire les capacités cognitives humaines au sein des machines, leur permettant ainsi de percevoir, raisonner, apprendre et s'adapter de manière autonome à des tâches complexes. Pour atteindre cet objectif, ces systèmes nécessitent une représentation adéquate des connaissances en jeu, ainsi que des mécanismes efficaces pour exploiter ces connaissances ou pour raisonner à partir d'elles.

Le Raisonnement à Partir de Cas (RàPC) est une technique appliquée dans divers domaines, tels que la médecine, l'ingénierie, la finance, la gestion des connaissances et le support client. En permettant aux systèmes informatiques d'apprendre à partir d'expériences passées et d'adapter leurs solutions à des situations spécifiques, le RàPC offre une approche flexible et efficace pour résoudre des problèmes complexes dans des environnements dynamiques et en constante évolution.

Le RàPC est une méthode relativement récente de résolution de problèmes, qui consiste à aborder de nouveaux défis en s'appuyant sur des solutions apportées à des problèmes similaires rencontrés précédemment (Aamodt et Plaza, 1994). Contrairement à d'autres méthodes qui reposent sur des règles explicites ou des modèles formels, le RàPC utilise des expériences passées, conservées sous forme de "cas", pour orienter la résolution des problèmes actuels.

Dans le domaine médical, la résolution de problèmes complexes, le diagnostic des maladies et la prise de décisions cliniques s'inspirent des pratiques des professionnels de santé qui tirent parti de leurs expériences passées pour évaluer et traiter les patients, intégrant ces connaissances dans un processus systématique et reproductible.

Au cœur du RàPC se trouve une base de données contenant des cas antérieurs, qui incluent des informations sur les symptômes, diagnostics, traitements et résultats pour diverses conditions médicales. Lorsqu'un nouveau cas est soumis, le système de RàPC cherche des similitudes avec des cas antérieurs et adapte les solutions passées aux besoins du patient actuel.

Dans cette thèse, notre travail se concentre sur le raisonnement à partir de cas, en particulier sur les phases de remémoration et d’adaptation. Nous proposons une modélisation de la remémoration dans le système RàPC, reposant sur l’utilisation de la forêt aléatoire pour classifier les cas stockés dans la base de données en fonction de leur pertinence par rapport au problème actuel, ce qui permet une remémoration plus efficace des cas similaires. Ensuite, nous développons un algorithme d’adaptation visant à adapter les cas remémorés en solutions plus robustes et généralisables.

Motivations et objectifs

Étant donné que les travaux dans cette thématique sont encore très limités dans le domaine médical, et que l'hybridation du Raisonnement à Partir de Cas (RàPC) avec d'autres techniques d'intelligence artificielle (IA) peut améliorer la précision et la performance du système RàPC tout en résolvant facilement des problèmes complexes, cela nous a motivés à proposer cette thématique.

Dans le cadre de cette thèse, l'objectif principal est de concevoir une approche hybride basée sur le raisonnement à partir de cas et les forêts aléatoires pour aider les experts du domaine médical à prendre des décisions éclairées. Les objectifs spécifiques de notre travail se résument comme suit :

- Modélisation de la base de cas à l'aide des forêts aléatoires.
- Intégration de techniques de sélection de caractéristiques sur les attributs des cas.
- Amélioration des performances du système RàPC à travers :
  - Une modélisation de la phase de remémoration par l'intégration de forêts aléatoires optimisées.
  - Une adaptation facile à interpréter afin d'améliorer le processus de remémoration.
- Réalisation d'une étude comparative entre l'approche proposée et d'autres modèles existants.

Contributions

Les travaux effectués dans le cadre de cette thèse tentent de répondre aux objectifs cités ci-dessous en vue de réaliser une approche hybride qui permet de classer les données médicales pour aider les experts de prendre les décisions précieuses.

Nos contributions portent sur deux volets majeurs :

La première contribution consiste à la réalisation d'une approche hybride intégrant des forêts aléatoires classiques ainsi que des forêts aléatoires modifiées par deux façons différentes : la première méthode utilise les caractéristiques pertinentes sélectionnées manuellement selon un médecin spécialiste, et la deuxième méthode, nous avons pondéré les caractéristiques par un poids donné. Nous avons également prié en compte l'importance des techniques de sélection de caractéristique manuel lors de la classification mais engendrant un cout en temps. Pour pailler à cet inconvénient, nous proposons dans la deuxième contribution.

Cette derniére est une approche hybride qui combine entre les techniques de sélection des caractéristiques automatiques et une modélisation d'une remémoration guidée par l'adaptation qui s'appuie sur des forêts aléatoires améliorée afin de réduire le cout en temps en construisant une approche plus efficace.

Organisation du manuscrit

Cette thèse est divisée en deux parties principales : la première expose les concepts de base et l'état de l'art, tandis que la seconde partie présente nos contributions majeures. La thèse débute par une introduction générale et se conclut par une conclusion générale.

Partie 1 : Cette section est structurée en trois chapitres :

- **Chapitre 1 : Raisonnement à Partir de Cas (RàPC)**

Ce chapitre introduit les principes fondamentaux du RàPC. Nous y présentons plusieurs définitions, les concepts clés de cette approche, ainsi que les techniques d'organisation de la base de cas. Nous offrons également un état de l'art des systèmes médicaux basés sur le RàPC et discutons des différences existant entre les diverses phases du processus RàPC.

- **Chapitre 2 : Forêts aléatoires (RF)**

Ce chapitre aborde les méthodes d'ensembles, en se concentrant sur les forêts aléatoires, le principe d'élagage, et les techniques employées pour éviter le sur-apprentissage dans les arbres de décision d'une forêt aléatoire. Un résumé des travaux médicaux basés sur les forêts aléatoires est également présenté.

- **Chapitre 3 : Techniques de sélection des caractéristiques**

Dans ce chapitre, nous définissons les différents concepts liés à la sélection des caractéristiques. Nous détaillons les étapes du processus de sélection, suivies par un état de l'art des méthodes de sélection des caractéristiques utilisées dans cette thèse.

Partie 2 : Cette partie est composée d'un seul chapitre :

- **Chapitre 4 : Contributions**

Ce chapitre présente les principales contributions de notre travail :

Contribution 1 : Modélisation de la remémoration d'un RàPC à l'aide d'une forêt aléatoire optimisée et de trois techniques de sélection des caractéristiques.

Contribution 2 : Comparaison entre un système hybride CBR-RF et CBR-DT pour la classification des données médicales.

Contribution 3 : Proposition d'une forêt aléatoire améliorée basée sur la sélection et la pondération des caractéristiques pour la remémoration des cas dans le système RàPC.

Contribution 4 : Développement d'une nouvelle forêt aléatoire optimisée pour la classification des données médicales, en utilisant la corrélation de Pearson et le meilleur nombre d'arbres.

Enfin, nous concluons cette thèse par une conclusion générale où nous dressons un bilan de notre travail et proposons des perspectives pour poursuivre cette recherche.

Chapitre 01

Raisonnement à Partir de ′Cas (RàPC)

Chapitre 01

Raisonnement à Partir de Cas (RàPC)

1 Introduction

Notre expérience nous permet naturellement de faire face aux défis de la vie quotidienne. L'intelligence artificielle (IA), qui intègre des capacités cognitives telles que la mémorisation de cas antérieurs, est connue sous le nom de Raisonnement à Partir de Cas (RàPC) (Althoff, 2011).

Le RàPC est une méthode de résolution de problèmes qui se distingue fondamentalement des autres approches couramment utilisées en IA. Cette méthode imite le processus de pensée humaine en se basant sur la recherche et l'adaptation de cas antérieurs. Un système de RàPC s'appuie sur une ou plusieurs solutions précédemment appliquées à des cas similaires pour résoudre de nouveaux problèmes (Aamodt et Plaza, 1994).

En raison de son approche de raisonnement qui s'aligne sur le raisonnement humain, le RàPC est l'une des techniques d'IA les plus répandues dans le domaine médical. Cette méthode consiste à résoudre un nouveau problème en utilisant l'analogie avec des problèmes passés, également appelés « cas sources ». Ces cas fournissent des informations précieuses pour aborder de nouveaux défis, désignés comme « cas cibles ». Les cas sources sont stockés dans une mémoire appelée « Base de Cas (BC) ».

Ce chapitre est structuré en cinq sections. Nous commencerons par présenter les définitions du Raisonnement à Partir de Cas. La deuxième section est dédiée à la description du cycle de vie du RàPC, en détaillant chacune de ses étapes. La troisième section explore les concepts fondamentaux du RàPC. Les dernières sections abordent les mesures de similarité utilisées dans cette approche, ainsi que quelques travaux médicaux basés sur le RàPC cités dans la littérature.

2 Raisonnement à partir de cas

Le Raisonnement à Partir de Cas (RàPC) a été introduit pour la première fois par Roger Schank et ses étudiants à l'Université de Yale au début des années 1980. Ce mécanisme présente une grande similitude avec les méthodes de diagnostic employées par les experts médicaux. Le RàPC est un exemple précieux de systèmes d'aide à la décision (Bichindaritz et Montani, 2011).

Il s'agit d'une méthodologie de raisonnement qui simule le raisonnement humain en utilisant les expériences passées pour résoudre de nouveaux problèmes (Kolonder, 1993). Le RàPC est un paradigme qui consiste à résoudre un problème nouveau, appelé cas cible, en s'appuyant sur un ensemble de problèmes déjà résolus (Kolonder, 1993). Un cas source fait référence à des problèmes similaires rencontrés et résolus dans le passé (Al Sun, 2012).

Les problèmes résolus sont stockés dans une base de cas, qui est un ensemble de cas du système (cas sources). Ces cas peuvent être fournis soit par l'utilisateur, soit enrichis automatiquement par le système. Selon (David, 2003), le raisonnement à partir de cas est un paradigme d'intelligence artificielle permettant de résoudre de nouveaux problèmes en récupérant et en adaptant des solutions issues d'épisodes antérieurs de résolution de problèmes.

3 Cycle du RàPC

Le cycle de base du raisonnement à partir de cas passe par quatre phases (la remémoration, la réutilisation, la révision, et la mémorisation). La figure 1.1 illustre ce cycle tel qu'il a été proposé par (Amodt et Plaza ,1994).

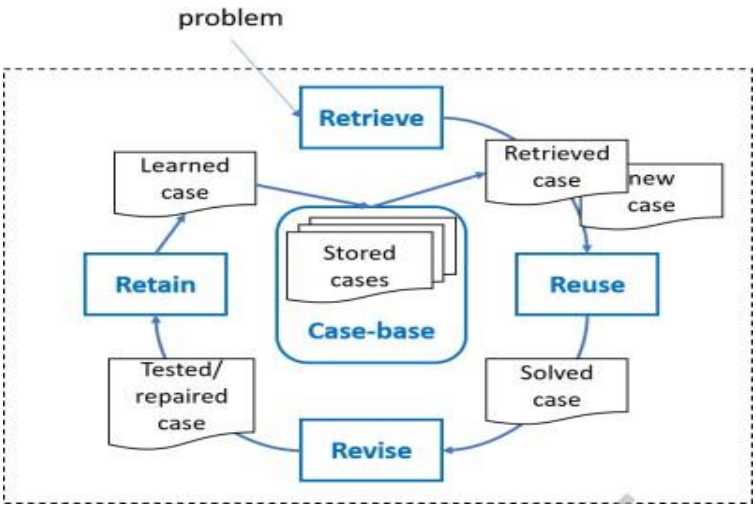


Figure 1.1. Cycle de RàPC (Aamodt et Plaza, 1994).

3.1 Phase de Remémoration (Retrieve)

La remémoration est une étape cruciale du processus de Raisonnement à Partir de Cas (RàPC), car elle est fondamentale pour le succès du système. Son objectif est de localiser ou d'extraire des cas pertinents de la base de cas afin de les utiliser comme références pour résoudre le problème cible. Les cas similaires sont identifiés à l'aide de diverses méthodes, telles que la distance euclidienne ou les voisins les plus proches.

3.2 Phase de réutilisation (Reuse)

Une fois qu'un cas pertinent est remémoré, il passe à l'étape de réutilisation, où la solution associée est testée pour évaluer son efficacité. Si la solution sélectionnée ne répond pas aux exigences du problème cible, elle doit être modifiée (adaptée). L'adaptation consiste à ajuster la solution du problème source pour qu'elle convienne au problème cible. La solution du problème source peut être conservée telle quelle ou modifiée pour proposer une nouvelle solution adaptée au problème cible. Ce processus d'adaptation vise à combler les écarts entre les deux cas, souvent à l'aide de mécanismes basés sur des règles. L'adaptation est généralement effectuée par des experts du domaine.

3.3 Phase de révision (Revision)

L'objectif de cette phase est d'évaluer et de réviser les solutions proposées par l'étape précédente en fonction de règles spécifiques au domaine d'application. Si la solution est appropriée, elle est intégrée comme une solution réussie. Sinon, le système ajuste la solution en conséquence.

3.4 Phase de Mémorisation (Retain)

La mémorisation est la dernière étape du cycle RàPC. Au cours de cette phase, le cas résolu avec succès est enregistré dans la base de cas pour une utilisation future. Le stockage d'un nouveau cas peut ne pas toujours être pertinent, mais il est essentiel d'éviter d'ajouter des cas redondants, car cela pourrait nuire aux performances du système.

Les principales étapes du processus RàPC sont : la remémoration, la réutilisation, la révision et la mémorisation. En complément, les structures de connaissance incluent le vocabulaire d'indexation, la base de cas, les métriques de similarité et les connaissances d'adaptation.

4 Concepts de base

4.1 Cas

Déterminer comment formuler un cas est la première étape du développement d'un système RàPC. Un cas dans une base de cas représente divers types de connaissances qui peuvent être stockés sous divers formats de représentation (Watson, 1997).

Un cas est l'ensemble formé par le problème et sa solution. Cas = {Problème, Solution (Pb)}

Un cas source comprend des caractéristiques (attributs) uniques pour décrire un problème, ces caractéristiques sont déterminées à l'avance par le concepteur du système.

Les deux principaux composants d'un cas sont la partie problème et la solution (Begum et Al, 2011) représente : une partie problème d'un cas et la partie solution de ce cas.

4.2 Base de cas

Une base de cas constitue une source essentielle de connaissances dans la méthodologie RàPC. Elle regroupe l'ensemble des cas résolus antérieurement et est structurée de manière à faciliter une recherche et une remémoration efficaces. La gestion de la mémoire et le processus de remémoration sont étroitement liés à l'organisation des cas résolus dans la mémoire (Nouaouria, 2013). L'organisation de la base de cas repose sur plusieurs types fondamentaux (Nouaouria, 2013) :

4.2.1 Organisation plate

Il s'agit d'une structure couramment employée dans les travaux RàPC, favorisant une mémorisation simple des cas et étant particulièrement adaptée à la gestion de petits ensembles de données. Les cas sont stockés dans un tableau organisé en lignes et colonnes, où chaque colonne représente un attribut spécifique. Cette approche présente l'avantage de permettre, lors de la phase de remémoration, l'évaluation de tous les cas disponibles en mémoire, garantissant ainsi une remémoration précise, sous réserve de la qualité de la fonction d'appariement.

4.2.2 Organisation hiérarchique

Lorsque la mémoire contient un grand nombre de cas, une organisation hiérarchique peut être utilisée pour faciliter la remémoration. Il existe principalement deux méthodes pour organiser hiérarchiquement les cas en mémoire : les réseaux de traits ou caractéristiques partagés et les arbres de discrimination.

5 Mesure de similarité

La recherche de cas similaires est un aspect fondamental de la RàPC. Souvent, la mesure de similarité entre cas est simplifiée en une mesure de similitude entre problèmes. La complexité des données et leur représentation jouent un rôle crucial dans ce processus. Avant de rechercher des cas similaires, il est essentiel d'analyser le problème en question et d'identifier ses caractéristiques.

Les mesures de similitude visent à trouver un cas dans la base de données qui est similaire au problème actuel, afin qu'il puisse être adapté au nouveau contexte. Ces mesures cherchent des correspondances entre les descripteurs des cas cibles et ceux des cas de la base lors de la phase de remémoration. Le calcul de la similarité ou de la dissimilarité peut être réalisé de différentes manières, la distance de Minkowski étant l'une des plus couramment utilisées.

5.1 Plus proches voisins (KPPV)

La technique des K plus proches voisins (KPPV), connue en anglais sous le nom de K-Nearest Neighbors (KNN) (Weiss et Kulikowski, 1991), consiste à identifier les points d'un ensemble qui sont les plus similaires ou les plus proches d'un point donné, en utilisant une mesure de distance. Pour résoudre ce problème, on calcule les distances entre chaque point de l'ensemble et le point à tester, puis on sélectionne les points ayant les distances minimales.

5.2 Méthodes inductives

Pour améliorer la remémoration, les méthodes inductives s'appuient sur l'organisation structurelle de la base de cas sous la forme d'un arbre de décision. Ces techniques segmentent les clusters de cas, en déterminant quels descripteurs sont capables de différencier les cas afin de construire une structure d'arbre de décision. Cette organisation permet de structurer les cas dans la base de données de manière plus efficace.

6 Systèmes médicaux basés sur le raisonnement partir de cas (RàPC)

Le Raisonnement à Partir de Cas (RàPC) est une approche utilisée dans les systèmes médicaux pour résoudre de nouveaux problèmes en réutilisant les expériences des cas précédents stockés dans une base de cas. Le système associe les nouveaux cas à des cas antérieurs similaires en utilisant des calculs de similarité et récupère les informations pertinentes pour fournir un diagnostic ou une réponse au nouveau cas. Il existe plusieurs travaux de raisonnement à partir de cas qui ont basé sur la remémoration dans le domaine médical. Nous citons ci-dessous quelques travaux.

Tableau 1.1. Aperçu de la littérature des travaux basés sur le système RàPC dans le domaine médical (Tarchoune et al, 2023)

Auteur(s), Année	**Technique(s) utilisée(s)**	**Zone de Candidature**	**Résultat**	**Taille dela base de cas**	**Phase implémentée**
(Perner, 1999)	RàPC	Analyses images medicals	-Le système obtenu de bons résultats sur l'analyse d'image -La performance r=0.85 c'est un très bon résultat	9 cas	Remémoration
(Montani et al, 2000)	RàPC Plus proche voisin	Diabète	-Intégré dans la gestion télématique du diabète insuline dépendant (T-IDDM)	147 cas réels.	Remémoration
(Marling et White house, 2001)	RàPC Rule-based reasoning	Alzheimer	-Système aide à la décision(SAD) pour la planification des soins continus des patients	Université Alzheimer Center	Remémoration
(Golobardes et al, 2002)	RàPC AG	Cancer du sein	-Améliorer la précision des prédictions du système. -une bonne fiabilité: convertir les cas mal classés en cas nonclassés	216 images	Remémoration

(Montani et al, 2003)	RàPC Rule-based reasoning	Diabetes de type 1	-Le système capable de détecter des situations complexes et de lestraiter de manière plus robuste	26 cas réels	Réutilisation (Adaptation)
(Vorobievaet Schmidt, 2003)	RàPC	Endocrinologie	-Le développement d'une méthodologie générale pour le problème d'adaptation des systèmes médicaux	/	Réutilisation (Adaptation)
(Perner et al, 2004)	RàPC Image processing	Reconnaissance des spores des champignons aéroportés	Développement d'une méthode pour identifier les spores dans une image microscopique numérique	La collection de champignons d'universitéde Jene Allemagne	Remémoration
(Quellec et al, 2008)	RàPC Arbre de décision	Générale	-Precision: (DRD) 81.8%. (DDSM) 84.8%.	-la rétinopathie diabétique(DRD) -mammographie de dépistage (DDSM).	Remémoration
(Houeland, 2011)	RàPC Arbre de décision	Cancer	-RDT permet de définirla similitude entre deux cas -les résultats ont été évalués dans le domaine des soins palliatifs du Cancer	-dossiers de patients dans le domaine des soins palliatifs	Remémoration
(Chattopadhyay et al, 2013)	RàPC KNN	Générale	-Son niveau est acceptable -L'application est large (générale)	53 cas	Remémoration
(Blanco et al, 2013)	RàPC	Générale	-Synthèse des articles publiés entre 2008-2011 dans la santé	1018 références	Réutilisation (Adaptation)
(Leal et al, 2013)	RàPC	Systèmes de surveillance continue du glucose	-Outil potentiel pour distinguer les mesures thérapeutiquement correctes et incorrectes dans les systèmes de surveillance continue du glucose(SGC)	22 cas Unité de soin intensif (USI)	Remémoration
(Sharaf-El-Deen et al, 2014)	RàPC Rule- based reasoning	Cancer du sein et la maladie de la thyroïde	-L'approche proposée augmente la précision de diagnostic des systèmes RàPC -Fournit une précision fiable par rapport aux systèmes actuels du diagnostic	215 cas Maladie thyroïdienne (UCI)	Réutilisation (Adaptation)
(Tyagi et Singh, 2015)	RàPC	Asthme	-Service de soins de l'asthme(ACS) utile d'établir un plandesoins automatique pour le patient -Améliorer la qualité de vie des patients	Service de soins de l'asthme	Remémoration

(Chakraborty et al, 2015)	RàPC	Cholera	-Le système CEDS aide à identifier le choléra -Minimiser les erreurs de déviation qui se sont avérées être une cause notable d'erreurs médicales	/	Remémoration
(Yin et al, 2015)	RàPC	Mal de tète	-CDSS RàPC avec un degré élevé de précision et obtenait de meilleurs résultats que le CDSS basé sur des lignes directrices -Performance diagnostique plus élevé pour les MP et les PTTH	676 cas cliniques	Remémoration
(Khussainova et al, 2015)	RàPC Clustering	Radiothérapie (cancer du cerveau)	-Le système à un taux de réussite plus élevé	86 cas Real	Remémoration
(Tahmasebian et al, 2016)	RàPC système flou	Rénales chronique	-L'utilisation de la méthode floue pour mesurer la similarité peut conduire à une plus grande flexibilité par rapport aux autres méthodes -outil puissant sur le lien de soin	100 cas Dossiers médicaux	Remémoration
(Nasiriet Fathi, 2017)	RàPC	Démence	-Classé et mis à jour par les soignants et lese xpertsdu domaine	37 Livres sur la démence	Remémoration
(Choudhury et Begum, 2017)	RàPC logique floue	Générale	-Permet de mettre au point des systèmes hybrides efficaces	/	Remémoration
(Ramos et al, 2017)	RàPC Gradient Boosting	Lung Cancer	-Apprendre au fil du temps. -Adaptabilité, -Interopérabilité de solutions.	/	Réutilisation
(Brown et al, 2018)	RàPC	Diabète de type 1	-Améliore l'étape de recherche -Meilleure recherche de cas par rapport à l'approche traditionnelle	/	Remémoration

(Bentaiba et al, 2018)	RàPC Randomisation	Gravitédela masse de la mammographie	-Améliorer l'efficacité de la résolution de la RàPC -Système plus rapide	100 cas	Remémoration
(Lei et Yin, 2019)	RàPC Mathématique floue	Générale	-La faisabilité et la validité de la méthode proposée -Méthode efficace	180 cas Organisation des dossiers	Remémoration
(Saadi et Henni, 2019)	RàPC Arbrede décision	Maladie des verrues	-Améliorer la performance de la phase de remémoration	Cas réels	Remémoration
(Bentaiba et al, 2020)	RàPC Randomisation	Masse mammographie et les maladies Des thyroïdes	-Améliorer la précision de la résolution -accélère la remémoration des cas. -Comparer avec d'autres méthodes de classification (un bon concurrent)	/	Remémoration
(Feuillâtre et al, 2020)	RàPC K-NN Arbre de décision	Cardiopathie valvulaire	-Améliorer la performance de la remémoration et là Réutilisation de cas	138 cas	Remémoration
(Bach et Mork, 2021)	RàPC Modélisation des mesures de similarités avec 8 classificateurs	Diabète	-Montré que les mesures de similarité sont responsables au développement des systèmes RàPC - la précision la plus performante est 78%	768 cas	Remémoration
(Gu et al, 2020)	RàPC XGBoost	Cancer du sein	-L'interopérabilité améliorelaconfiancedes médecins dans le système. -la précision la plus performante est 91%	217 cas	Remémoration
(De et Chakraborty, 2021)	RàPC	Anémie	-Réduire le temps de formation -Le système plus performant avec précision de 92%	200 cas	Remémoration
(Martinez et al, 2021)	RàPC KNN	COVID-12	-Effective proposal	/	Remémoration

(Wang et al, 2022)	RàPC KNN	/	-La précision est supérieure à 81%. -RàPC peut être appliqué pour résoudre des Problèmes compliqués dans la pratique	11 datasets	Remémoration
(Mustafa et al, 2023)	RàPC Rule-based reasoning	Cancer du sein Maladies cardiaques Diabète Activité et de biométrie des smartphones et des montres intelligentes WISDM	Les deux systèmes présentés dans cet article ont obtenu des résultats compétitifs et fournissent des diagnostics similaires à ceux réalisés par des experts humains.	4 datasets	Remémoration
(Admass et Munaye, 2024)	RàPC Rule-based reasoning	Maladies de la mangue	RàPC et RBR améliorent considérablement la précision et l'efficacité du système	données du monde réel collectées auprès de l'Institut éthiopien de recherche agricole et de l'Agence nationale de météorologie	Remémoration

6.1 Discussions et Résultats

Le Tableau 1.1 présente par ordre chronologique, certains des systèmes RàPC développés dans le domaine médical au fil des années. Il classe également ces systèmes en fonction des techniques utilisées, de la taille de la base de données et tente de déterminer la phase dans laquelle chaque système RàPC se situe.

Les systèmes RàPC en médecine se concentrent principalement sur le diagnostic, la classification et la planification. L'utilisation de systèmes RàPC complets dans ce domaine est relativement rare. Notre enquête montre que la majorité des systèmes médicaux RàPC réussis sont construits autour d'une combinaison de RàPC et d'autres méthodes d'intelligence artificielle (IA). Bien que le RàPC soit précieux en raison de son raisonnement similaire à celui de l'humain, il présente des limites. Par conséquent, la plupart des systèmes médicaux RàPC sont devenus hybrides pour surmonter ces contraintes.

Depuis la fin du siècle dernier, des systèmes hybrides ont émergé pour les systèmes RàPC médicaux. Selon (Sun et al, 2008), un système RàPC hybride peut être considéré sous cinq perspectives différentes. Ces systèmes RàPC hybrides sont souvent associés à divers algorithmes, tels que le raisonnement basé sur les règles, la logique floue, les arbres de décision, la randomisation, et les algorithmes génétiques, entre autres.

Le Tableau 1.1 montre que presque tous les systèmes sont hybrides (24 systèmes), bien que

quelques-uns (10 systèmes) reposent uniquement sur le RàPC. Cela démontre que l'hybridation des systèmes RàPC offre des opportunités prometteuses pour améliorer ces systèmes. Parmi les applications de ces systèmes, on trouve la sélection des caractéristiques, l'extraction des caractéristiques, la pondération des caractéristiques, et l'adaptation. Les systèmes RàPC hybrides récents utilisent des méthodes d'exploration de données telles que la logique floue, les réseaux de neurones, les arbres de décision, et les forêts aléatoires pour le prétraitement des données ou comme mesures de similarité.

La phase de remémoration dans un système RàPC est cruciale, car elle repose sur la recherche de cas similaires au problème donné. La similarité est un concept central du RàPC, utilisé non seulement pour la recherche de cas, mais aussi pour l'adaptation des cas. Parmi les techniques employées pour mesurer la similarité dans les systèmes RàPC médicaux figurent le plus proche voisin, la distance euclidienne, et les algorithmes génétiques.

Les combinaisons de ces techniques ont été rapportées pour définir les différentes étapes du RàPC. L'étude a révélé que la plupart des systèmes parviennent à bien définir la phase de remémoration, qui est un composant clé du RàPC. Parmi les 33 systèmes examinés, 28 systèmes ont réalisé cette phase (voir Figure 1.3) et fonctionnent principalement comme des systèmes de remémoration.

L'adaptation du cas peut être une étape nécessaire si la solution ne répond pas aux besoins du problème. De nombreux systèmes médicaux RàPC évitent l'étape d'adaptation automatique en raison de sa complexité par rapport aux autres phases, du grand nombre de caractéristiques, et des changements rapides dans les connaissances médicales. Toutefois, l'adaptation est souvent une question difficile dans ce domaine et est souvent effectuée manuellement par des experts médicaux.

Selon (Smiti et Elouedi, 2014), le succès d'un système de raisonnement basé sur les cas dépend de la qualité des données des cas et de la rapidité du processus de remémoration, qui peut être coûteux en temps, surtout lorsque le nombre de cas est élevé. Pour garantir cette qualité, il est nécessaire de maintenir le contenu de la base de cas. Une base de cas statique et non évolutive limite la précision du RàPC dans la résolution des problèmes. Par conséquent, la maintenance de la bibliothèque de cas, c'est-à-dire l'augmentation des connaissances et de l'expérience dans le RàPC (les cas résolus dans la base de cas), peut affecter le temps de résolution. La construction de la base de cas est une forme de raisonnement basé sur la similarité. L'idée centrale est que les relations de similarité ne sont pas seulement utilisées pour la recherche de cas, mais aussi pour créer la base de cas.

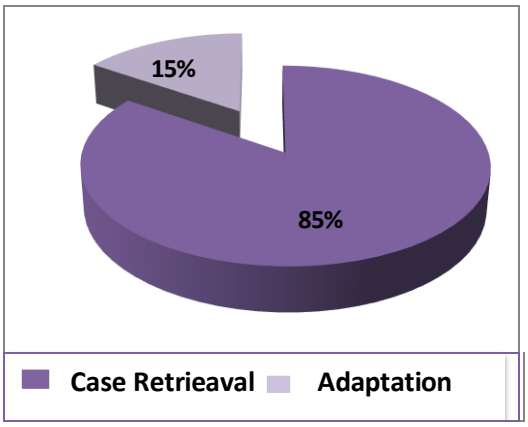


Figure 1.2. Pourcentage des systèmes en termes de phase implémentée (Tarchoune et al, 2023)

Vers une approche hybride

L'étude a également révélé que l'hybridation de la RàPC avec d'autres techniques d'IA améliore la précision et la performance des systèmes RàPC, permettant de résoudre plus efficacement des problèmes complexes grâce à plusieurs motivations de développement :

- Exploiter les avantages des deux approches.
- Combiner des méthodes analytiques et heuristiques complémentaires.
- Améliorer potentiellement les performances du système en intégrant toutes les connaissances disponibles des deux sources.

À cet égard, nous proposons d'intégrer les forêts aléatoires pour la modélisation de la base de cas, ainsi que comme mesure de similarité dans la première phase du système RàPC.

7 Conclusion

Le raisonnement à partir de cas est une approche largement adoptée par les cliniciens, car elle reflète le mode de pensée humain. Nous avons présenté les principes fondamentaux d'un système de raisonnement à partir de cas, en détaillant les différentes phases du cycle qui permettent la manipulation des mécanismes de connaissance. Cette présentation a révélé que ces phases sont interdépendantes, en particulier les phases de remémoration et d'adaptation, que nous avons examinées en détail au chapitre 4.

Ce premier chapitre a établi les bases nécessaires pour comparer les différents systèmes de RàPC appliqués au diagnostic médical. Il est important de noter que la mesure de similarité est au cœur du RàPC, car elle permet d'identifier des correspondances entre les descripteurs des cas cibles et des cas sources lors de la phase de remémoration. C'est pourquoi nous avons sélectionné les forêts aléatoires pour la modélisation de la base de cas et comme mesure de similarité dans la phase de remémoration, sujet que nous abordons au chapitre 2.

Chapitre 02

Forêts Aléatoires (RF)

Chapitre 02

Forêts Aléatoires (RF)

1 Introduction

L'algorithme des forêts aléatoires est une des méthodes statistiques les plus utilisées et l'une des plus efficaces pour effectuer des prédictions, cette méthode permet également de trier les valeurs en fonction de leur importance pour prédire la valeur d'intérêt souhaitée.

Dans un premier temps, nous avons présenté les méthodes d'ensembles, la méthode des forêts aléatoires a été inspirée par les méthodes d'ensemble dans la deuxième section. Dans la troisième section, nous abordons le principe d'élagage et ces techniques, qui sont utilisé pour éviter le sur-apprentissage des arbres de décisions dans une forêt aléatoire. Ensuite, dans la quatrième section, nous avons présenté un résumé des différents travaux médicaux qui sont basées sur les forêts aléatoires dans la littérature. Finalement, dans la dernière section, la solution que nous avons proposée est une hybridation entre le raisonnement à partir de cas et les forêts aléatoires, avec un résumé des différents chercheurs qui sont entamé cette hybridation.

2 Méthodes d'ensemble

En règle générale, les méthodes d'ensemble se basent sur la création d'une série de prédicteurs pour ensuite combiner toutes leurs prédictions. L'agrégation est une méthode de classification qui nécessite un vote majoritaire parmi les classes proposées par les prédicteurs. La construction aléatoire d'une famille de modèle « Boosting » et adaptative, déterministe ou aléatoire d'une famille de modèle « Bagging » sont les bases des méthodes d'ensemble.

2.1 Bagging

Principe générale : agréger une collection de classificateurs faibles pour obtenir un meilleur classificateur. En général pour la classification, agrégation par vote majoritaire.

La méthode du Bagging a été introduite par (Breiman, 1996). Le mot Bagging est la contraction des mots Bootstrap et Aggregating. Le Bagging repose sur la méthode Bootstrap.

1 : Bootstrap: On construit q échantillon $\mathcal{L}$n (1), …, $\mathcal{L}$n(q) à partir d'un seul échantillon de départ $\mathcal{L}$n= (xi, yi) i=1, …, n.

Un échantillon bootstrap $\mathcal{L}$n (1) est obtenu par tirage avec remise de n éléments parmi $\mathcal{L}$n ou chaque observation (xi, yi) a une probabilité

1 : d'être tiré à chaque tirage.

2 : Chaque échantillon bootstrap $\mathcal{L}$n (1) sert à construire un classifieur.

3 : On agrège cette collection de classificateur.

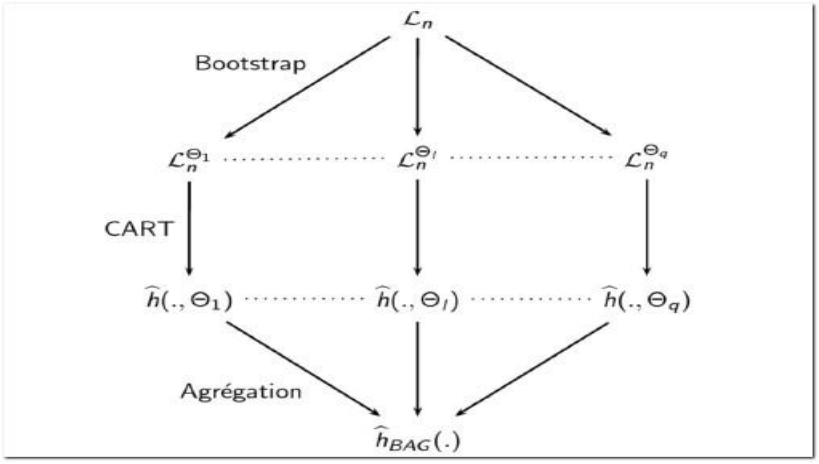


Figure 2.1. Principe du Bagging (Breiman, 1996)

2.2 Bootstrap

Principe générale : Le rééchantillonnage statistique est couramment utilisé pour estimer des grandeurs ou des propriétés statistiques.

Le bootstrap repose sur l'idée de créer plusieurs ensembles de données en procédant à des tirages aléatoires avec remise à partir de l'ensemble de données original. Chaque classifieur élémentaire est alors formé sur un de ces échantillons bootstrap, ce qui garantit que chaque classifieur est entraîné sur un ensemble d'apprentissage distinct. En agrégeant les prédictions de ces classifieurs, on obtient un prédicteur global plus performant.

2.3 Forêts aléatoires

La forêt aléatoire est une méthode statistique non paramétrique reconnue pour ses performances exceptionnelles, introduite par Breiman en 2001.

Principe : L'objectif est de réduire la corrélation entre les arbres issus du bagging.

Contrairement au bagging classique, la technique des forêts aléatoires introduit un critère de décorrélation entre les arbres en sélectionnant aléatoirement un sous-ensemble de variables à chaque niveau de décision pour déterminer le meilleur nœud de l'arbre. Ce choix aléatoire des variables vise à diminuer la corrélation entre les arbres sans augmenter excessivement leur variance.
Breiman a amélioré le bagging, notamment pour les modèles CART (arbres binaires). Lors de la construction des arbres, l'algorithme CART choisit la meilleure partition en fonction de l'indice de Gini à chaque nœud. Cependant, la sélection de la partition optimale se fait uniquement sur un sous-ensemble d'attributs, préalablement défini de manière aléatoire à partir de l'espace original des attributs. La prédiction globale de la forêt aléatoire est obtenue par la majorité des votes des arbres individuels. Cette approche fait partie des familles les plus étendues des forêts aléatoires, telles que définies par Breiman.

2.3.1 Définition des forêts aléatoires

Soit $\{\hat{h}(., \Theta 1), \ldots, \hat{h}(., \Theta r)\}$ une collection de prédicteurs par arbre, où $(\Theta 1, \ldots, \Theta r)$ est une suite de variables aléatoires, c'est-à-dire, indépendante de l'échantillon d'apprentissage $\mathcal{L}o$. Le prédicteur des forêts aléatoires est obtenu par agrégation de cette collection de prédicteur.

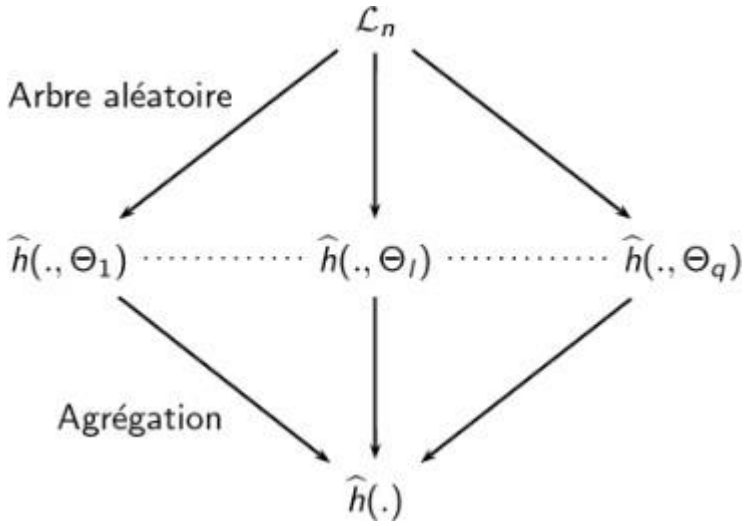


Figure 2.2. Principe des Forêts Aléatoires (Breiman, 2001)

2.3.2 Algorithme des forêts aléatoires

Les forêts aléatoires présentent plusieurs caractéristiques qui expliquent leur succès remarquable dans les tâches de classification. Contrairement à de nombreux autres algorithmes d'apprentissage automatique, les forêts aléatoires ne nécessitent pas un réglage minutieux des paramètres pour atteindre des performances prédictives. Les principaux paramètres influents sont : le nombre de variables tirées aléatoirement à chaque nœud (mtry), le nombre total d'arbres (M), la profondeur maximale des arbres (tree_depth), et le nombre minimal d'observations dans chaque feuille terminale (min_node_size).

Algorithme 2.1 : Forêts Aléatoires (Breiman, 2001)

Entrée: nombre_d'arbres, nombre_de_classifieurs_faibles, profondeur_maximale.

Sortie: Forét aléatoire.

Pour i de 1 à N

1. Construire un échantillon boostrap des données
2. Construire un arbre CART maximal à partir de cet échantillon boostrap, tel que:
3. A chaque nœud, on sélectionne la meilleure découpe sur la base de K variables choisies aléatoirement parmi les P variables d'entré.

Fin pour

Les forêts aléatoires se distinguent par leur stabilité, leur résistance au sur-apprentissage, et leur efficacité même dans des espaces de grande dimension. Ces propriétés sont en grande partie dues au principe de bagging, qui consiste à agréger un grand nombre d'arbres générés de manière aléatoire. La construction de chaque arbre étant indépendante des autres, il est possible de construire une forêt en parallèle, ce qui simplifie le calcul.

Un des principaux avantages des forêts aléatoires est leur capacité à éviter le sur-apprentissage, un problème fréquent dans les méthodes d'induction. Selon (Breiman, 2001), lorsque le nombre d'arbres dans la forêt augmente, le taux d'erreur en généralisation tend à converger vers une valeur limite. Cette valeur limite, une borne supérieure, peut être estimée en fonction des caractéristiques intrinsèques de la forêt. Cette propriété de convergence explique pourquoi les forêts aléatoires évitent le sur-apprentissage à mesure que le nombre d'arbres augmente, convergeant plutôt vers une erreur de classification hors échantillon (OOB) stable.

2.3.3 Erreur en dehors du boostrap

Le jeu de données de validation, également connu sous le nom de jeu de données hors sac (Out-of-Bag, OOB), comprend les exemples qui n'ont pas été inclus dans les échantillons bootstrap utilisés pour entraîner les arbres de la forêt aléatoire. Ce paramètre est essentiel pour effectuer une évaluation interne du classifieur et pour estimer l'importance des variables lors de la sélection des attributs.

3 Élagage

Les algorithmes de forêt aléatoire visent à améliorer la stabilité des modèles prédictifs en étendant les concepts des arbres de décision. Pour chaque arbre, des échantillons sont sélectionnés avec remplacement à partir des données. Certains échantillons peuvent ainsi apparaître plusieurs fois dans un même arbre. Un problème courant avec les arbres de décision est le sur-ajustement, qui se produit lorsque l'arbre devient trop complexe et s'adapte excessivement aux données d'entraînement. Une méthode simple pour éviter ce problème est de limiter la profondeur maximale de l'arbre, ce qui permet de contrôler la complexité du modèle.

Pour éviter le sur-apprentissage, on peut également élaguer les sous-arbres les moins significatifs. L'élagage consiste à supprimer des branches de l'arbre et à les remplacer par des feuilles, à condition que cette opération réduise l'erreur de généralisation. Cette phase se déroule après la construction initiale de l'arbre.

Lors de l'élagage, on commence par vérifier si chaque sous-arbre peut être remplacé par une feuille sans augmenter l'erreur de prédiction. Si tous les nœuds d'un sous-arbre sont des feuilles et que l'erreur de la feuille de remplacement est inférieure à celle du sous-arbre, alors le sous-arbre est remplacé par cette feuille. Pour déterminer l'étiquette de la nouvelle feuille, on analyse les étiquettes des éléments des feuilles du sous-arbre et on choisit la valeur la plus fréquente. Cette procédure est répétée jusqu'à ce qu'il ne soit plus possible de remplacer un sous-arbre par une feuille sans dégrader les performances du modèle.

3.1 Pré-élagage

Le pré-élagage intervient au cours de la construction de l'arbre et sert de critère d'arrêt pour l'expansion de l'arbre. Il consiste à établir des conditions spécifiques qui interrompent le processus de construction lorsque certaines limites sont atteintes. Cette méthode empêche l'expansion excessive des branches lorsque les attributs restants ne sont plus utiles pour diviser les exemples. Sans cette interruption, les branches continueraient à croître en utilisant les attributs restants, même lorsque les exemples au niveau du nœud atteignent le seuil deviendraient similaires à ceux des nœuds terminaux. Dans les algorithmes tels que C4.5 et CART, si la mesure de segmentation atteint une valeur nulle ou infinie, un nœud terminal doit être créé. Il est souvent préférable de stopper l'expansion des branches plutôt que de continuer sans amélioration significative dans la classification.

La technique de pré-élagage est adoptée par l'algorithme CHAID (KASS, 1980), Consiste à imposer des règles d'arrêt lors du développement de l'arbre. C'est-à-dire nous allons décider ou nonde continuer à développer un certain nœud (l'élagage au début). Ce qui revient à fixer une condition d'arrêt pour bloquer la construction**.** Nous avons cherché tous les combinaisons possibles en utilisant des paramètres différentes de nombre d'arbre, nombre de classifieur, et profondeur maximal, et garder l'arbre avec le meilleur résultat (Algorithme 2.1).

Algorithme 2.2: Pré- élagage

Entrée: nombre_d'arbres, nombre_de_classifieurs_faibles, profondeur_maximale.

Sortie: T (pruné) avec la meilleure performance.

1. Création du nœud jusqu'à ce que la performance sur la validation diminue.
2. Pour chaque nœud, développez selon une condition d'arrêt.
3. Trouvez les combinaisons possibles en utilisant différents paramètres.
4. Réservation du nouvel arbre.

3.2 Post-élagage

Le post-élagage est une méthode couramment employée dans de nombreux algorithmes d'apprentissage automatique, et elle est appliquée une fois que l'algorithme d'expansion de l'arbre est terminé. Par exemple, l'algorithme C4.5 utilise une estimation de l'erreur réduite de l'arbre, qui est dérivée de l'erreur apparente observée au sein de l'arbre.

Le post-élagage est préconisé par KART (Breiman et al, 1984), consiste d'abord à construire récursivement un arbre en le laissant croître jusqu'à atteindre son maximum au risque de

sur-apprendre (Lounici et al, 2014). Le modèle (arbre) obtenu n'est pas optimal, il reflète presque fidèlement les exemples de la base d'entrainement. Ensuite, dans une seconde phase élague les éléments qui contribuent au sur-apprentissage, soit remplacer le sous arbre par une feuille, soit rassembler deux nœuds en un seul, selon un critère qui est calculées par la formule ci-dessous :

$$Critère\ (Tk)\ = \frac{\mathrm{MC\ (d,k) - MCT\ (d,k)}}{\mathrm{N(k)\ (Nt\ (d,k) - 1)}} \qquad (2.1)$$

d: le nœud de l'arbre, **k**: version de l'arbre élagué, **Tk** : l'arbre obtenu après l'élagage d'un sous arbre.

MC (d, k) : Ce terme représente le nombre d'individus mal classés dans l'ensemble d'apprentissage par le nœud ddd de l'arbre TkT_kTk, en supposant que ce nœud soit transformé en feuille.

MCT (d, k) : Il désigne le nombre d'individus mal classés dans l'ensemble d'apprentissage par les feuilles situées sous le nœud ddd de l'arbre TkT_kTk.

N(k) : Ce terme correspond au nombre total de feuilles dans l'arbre TkT_kTk.

Nt (d, k) : Il représente le nombre de feuilles dans le sous-arbre de TkT_kTk situé sous le nœud ddd.

Nœud à élaguer : Le nœud à élaguer est celui qui présente le critère le plus bas. Lorsque ce nœud est élagué, il devient une feuille avec la classe majoritaire (voir Algorithme 2.2).

Algorithme 2.3: Post-élagage

Entrée: M échantillons et N caractéristiques.

Sortie: T (pruné) avec la meilleure performance.

1. Construisez l'arbre complet en utilisant (C4.5, CHAID, ID3).
2. Juste que nous obtiendrons le critère le plus bas.
3. Pour chaque nœud, laissez tomber les feuilles de l'arbre.
4. Trouvez le nœud que nous allons élaguer selon le critère (TK, D).
5. Transformez le nœud en une feuille avec la classe majoritaire.
6. Appliquez le jeu de validation sur les sous-arbres trouvés.
7. L'arbre qui obtient le taux d'erreur minimal est l'arbre final.

4 Systèmes médicaux basés sur les forêts aléatoires

A ce jour plusieurs recherches sont intéressées aux forêts aléatoires, nous présentons de cette section l'état de l'art des améliorations des RF dans le domaine de la santé.

Tableau 2.1. Aperçu de la littérature des travaux basés sur les forêts aléatoires (RF) dans le domaine médical (Tarchoune et al, 2023)

Auteur, Année	**Technique(s) utilisée(s)**	**Zone de Candidature**	**Résultat**	**Base de cas**
(Akin Ozcift, 2011)	Forêt aléatoire	Cancer du côlon, Cancer de la leucémie, Cancer du sein et Cancer du poumon	-La stratégie est efficace -Comparer avec 15 classificateurs -Meilleure precision de classification	Cancer du côlon, Cancer de la leucémie, Cancer du sein et Cancer du poumon
(Nguyen et al, 2013)	Forêt aléatoire Sélection de caractéristique	Cancer du sein	-Plus efficace et intéressante -Précision est de **99.8%** -Système très utiles aux médecins	Wisconsin
(Farzana et Nooraini, 2013)	Forêt aléatoire	Cancer du sein	-Précision est de **72%** -Méthode performante	700 cas
(Cherry et al, 2014)	Forêt aléatoire SVM	Lymphadénopathie abdominale	-améliorer la performance	30 cas
(Mishra et Suhas, 2016)	Forêt aléatoire Sélection caractéristique	Lésions osseuses de la colonne vertébrale	-Précision est de **91%** -Améliore la précision de la détectiondeslésionsosseuses	79 cas
(Razak et al, 2016)	Forêt aléatoire	Classification des MIRNA (cancer)	-Précision élevé presque **100%** -Le système est capable d'identifier les marqueurs MIRNA responsable de la classification du cancer	399 cas
(Wu et al, 2017)	Forêt aléatoire	Tuberculose	-Précision est de **81%** -Comparer à d'autres méthodes et RF a la meilleure précision	Dossier médicaux
(Vijayakumari et Manikumaran, 2017)	Forêt aléatoire	Poumons	-Précision est de **88.4%** -L'algorithme proposé est performant	4 types de maladies pulmonaires
(Liu et al, 2017)	Forêt aléatoire Sélection des caractéristiques	Analysedes maladies complexes	-Améliore la précision de la classification	11dossiers médicaux

(Xu et al, 2017)	Forêt aléatoire	Cardiovasculaire (CVM)	-Testé à deux sources de données -CHDD précision de **91.6%** -PKU précision de **97%** -Plus performant.	-Base (CHDD) -Base (PKU)
(Boseet al, 2017)	Forêt aléatoire	Epilepsie	-RF est le plus efficace avec une précision de **98%.**	100 cas (EEG)
(Hasan et al, 2018)	Forêt aléatoire Sélection des caractéristiques KNN Arbre de décision (ID3) Bayes naïfs Régression logistique	Maladies cardiaque	Comparaison de performance des algorithmes de classification dans la prédiction des maladies cardiaque **Résultat** -La régression logistique a obtenu de meilleurs résultats avec Précision de **92.76%** -Et RF classé la deuxième avec précision de **88.12%**	303 cas
(Devi et al, 2018)	Forêt aléatoire	Donnée variée	-Précision est de **95%** -Permet une remémoration rapide des données -une classification de données supérieures en un temps minime	/
(Kuo et al, 2018)	Forêt aléatoire Machine à vecteurs de soutien Arbre de décision (C4.5) Régression logistique	Coûts de la fusion vertébrale	-RF avait la meilleure performance -Précision est de **84%** -RF utilisé pour prédire les couts médicaux	532 cas Base de donnée 2010-2013
(Mohapatra et mohanty, 2018)	Forêt aléatoire Sélection des caractéristiques	Arythmie	-Précision est de **96%** -Performant	UCI
(Benbelkacem et Atmani, 2019)	Forêt aléatoire	Diabète	Les RF ont comparé à d'autres méthodes d'apprentissage par machine **Résultat** RF est le plus efficace avec taux d'erreur de **0.21à 40 arbres**	Pima

(Javeed et al, 2019)	Forêt aléatoire	Insuffisance cardiaque	-Précision est de **93.3%** -Présente une plus faible complexité temporelle -Meilleure performance	Cleveland
(Wang et al, 2019)	Forêt aléatoire	Sommeil à mouvements oculaires rapides	-Précision **84%** -La performance de Classification a montré une tendance non monotone	45 cas
(Wang et al, 2019)	Forêt aléatoire Extraction des règles	Cancer du sein	-Le processus est plus interprétable et plus précis -Améliorer la performance de diagnostic du cancer	WDBC WOBC SEER Wisconsin
(VijiyaKumar et al, 2019)	Forêt aléatoire	Diabète	-Le taux de précision est supérieur -Précision est de **90%** -Donne la meilleure prédiction du diabète	UCI
(Anirudh et al, 2019)	Forêt aléatoire Arbrede décision	Diabète	Précision est de -**72%** pour arbre de décision -**76.5%** pour RF	Pima
(Vignesh et Revathy, 2019)	Forêt aléatoire	Tumeur	-Couteux en temps -Précision est de **94.34%**	L'imagerie médicale
(Proniewska et al, 2020)	Forêt aléatoire	Surveillance du sommeil	-Précision est de **89%** -Outil puissant pour la classification des événements respiratoires pendant le sommeil	Les signaux ECG et acoustiques
(Yu et al, 2020)	Forêt aléatoire	Plusieurs maladies	-contrôler la taille complexe de l'arbre et éviter le sur- ajustement. -Précision max est de **94%**	UCI
(Alam et al, 2020)	Forêt aléatoire	10 maladies	-Système performant -Précision max est de **97%**	UCI
(Li et al, 2020)	Forêt aléatoire	Colorectal Cancer	-Précision efficace	Etats unis et la chine DPGAN
(Raposo et al, 2020)	Forêt aléatoire	Résistance aux Médicaments de VIH	-Plus grande stabilité	Stanford HIVDrug Resistance HIVDB

(Kabiraj et al, 2020)	Forêt Aléatoire XGBoost	Cancer du sein	-Précision de RF est **74.73%** -Précision de XGBoost est **73.63%**	275 cas
(Hong Yang et al, 2020)	Forêt aléatoire	Carcinome à cellules claires du rein	-a plusieurs limites parce que le pronostic du carcinome à cellules clairesdu rein est associé à de multiples variations génétiques	368 cas
(Asadi et al, 2021)	Forêt aléatoire	Maladies cardiaques	- La méthode proposée est comparé à d'autres méthodes d'apprentissage par machine -Méthode performante avec précision de **88%.**	6 ensembles de données UCI
(Ono et Mitani, 2021)	Forêt Aléatoire XGBoost	Cancer du sein	-RF est le meilleur entermes du taux d'erreur moyen.	400 cas
(Yao et Li, 2022)	Forêt aléatoire Sélection des caractéristiques	Dépistage du cancer du col de l'utérus	-améliorer la performance	280 cas
(Shi et al, 2022)	Forêt aléatoire	Maladie du genou	Précision de classification est **75%**	100 cas
(Tie et al, 2022)	Forêt aléatoire DeepWalk	Metabolite- Disease	Prédiction fiable	3460 cas
(Pavithra et Geetha, 2022)	Forêt aléatoire Machine à vecteurs de support	Cancer du foie chronique	-Précision de RF est **98%** -Précision de SVM est **99%**	280 cases
(Lilhore et al, 2023)	Forêt aléatoire Machine à vecteurs de support Sélection des caractéristiques	Hépatite C	Les résultats expérimentaux prouvent l'importance de la sélection de caractéristiques pour obtenir une plus grande précision dans la recherche sur le VHC.avec une précision de 96.82%	UCI
(Yadav et al, 2024)	Forêt aléatoire + réseaux de neurones Arbre de décision+ les k-plus proches voisins	Santé mentale	précision de (DT + kNN) est 86,69 % précision de (RF + NN) est 93,54 %,	Mental_health
(Gangwar et al, 2024)	Forêt aléatoire KNN SVM	Maladies cardiaques	Forêt aléatoire a obtenu le taux	UCI

4.1 Discussions et Analyses

Dans cette section, nous analysons les systèmes médicaux qui utilisent les forêts aléatoires pour les problèmes de classification. Les forêts aléatoires se distinguent comme l'un des meilleurs classifieurs pour la prédiction de diverses maladies.

Le tableau 2.1 répertorie les systèmes en fonction de leurs taux de précision et de la taille des bases de données. La plupart des systèmes présentent des performances remarquables, avec des précisions variant de 72 % à près de 100 %. La littérature disponible indique que les forêts aléatoires offrent des niveaux de précision parmi les plus élevés comparés à d'autres modèles dans le domaine de la classification.

Les figures (2.3 et 2.4) dessinent une comparaison graphique entre la taille des ensembles de données utilisés dans certains des articles présentés et la précision de classification.

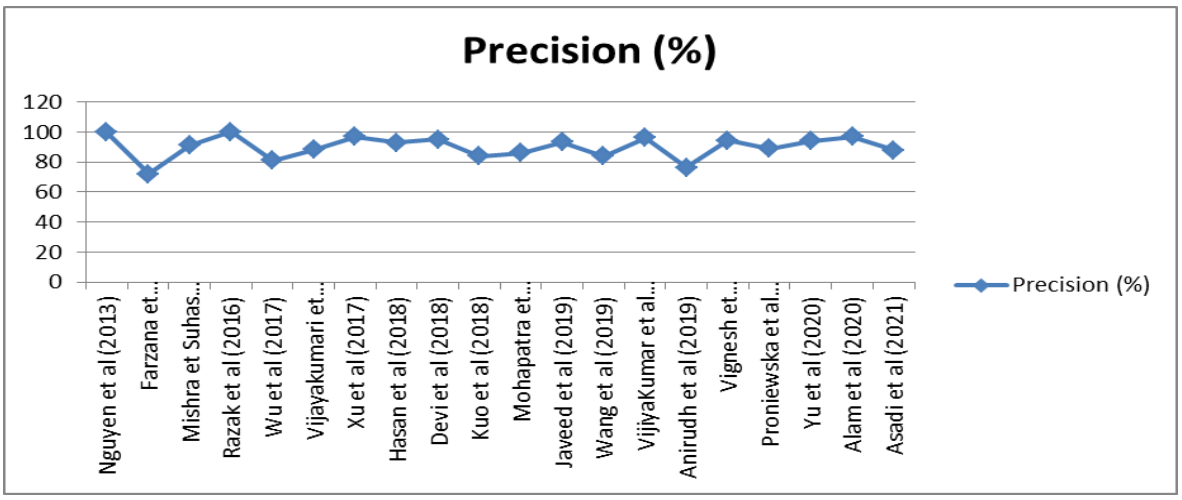


Figure 2.3. Taux de précision des systèmes RF (2011-2022) (Tarchoune et al, 2023)

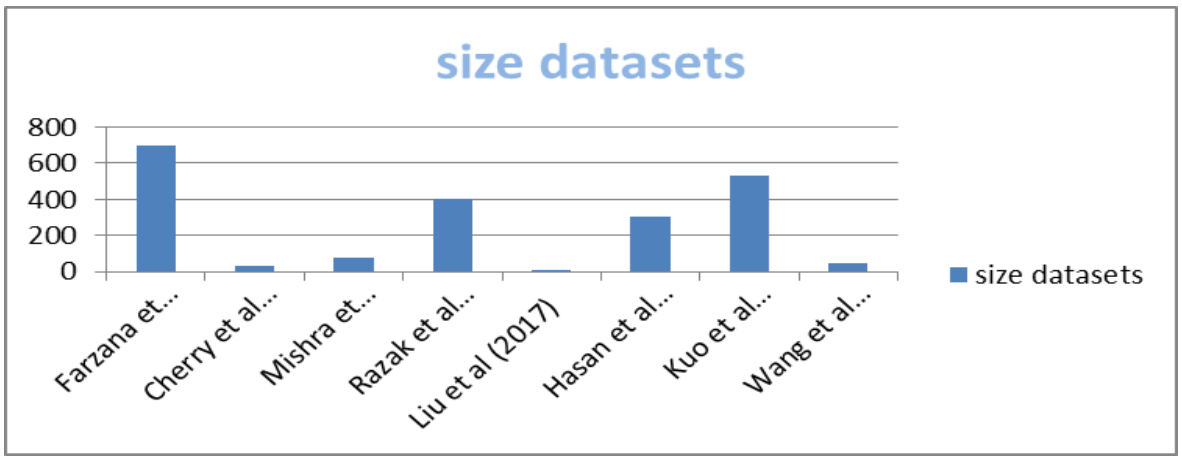


Figure 2.4. Taille des bases de données des systèmes traités (Tarchoune et al, 2023)

La forêt aléatoire a été utilisé avec succès dans différent taches notamment le développement de médicament, le cancer du sein, la Tuberculose, le diabète, les maladies cardiaques d'autre tel que les poumons, l'arythmie, surveillance du sommeil…etc.

Des recherche sont été effectuées en utilisant la forêt aléatoire comme classifieur et algorithme de sélection des caractéristiques pour le diagnostic médical, (Akin Ozcift, 2011) a utilisé le meilleur algorithme de RF de première recherche pour sélectionner les caractéristiques sur quatre ensemble de données médicales, le cancer de colon, le cancer du sein, le cancer de la leucémie et le cancer du poumon. Le modèle proposé avec la précision des caractéristiques extraies a été comparé à 15 classifieurs largement utilisés et formés avec toutes les caractéristiques et a montré une meilleure précision de classification.

Parmi les intérêts des RF, déterminer les prédicteurs les plus importants qui devraient être inclus dans un modèle performant, ceci peut être réalisé en effectuant les techniques de sélection de caractéristiques, dans laquelle peuvent être réduire la charge de données et réduire le temps de calcul.

L'extraction de caractéristiques devient une étape importante dans les récents systèmes médicaux à cause des données complexes ces derniers ont fait des précisions très élevés par rapport aux autres entre 91-99%. Ces travaux montrent l'efficacité des méthodes de sélection de caractéristiques dans le cadre de RF, la sélection de caractéristiques est souvent une partie nécessaire du développement du modèle de prédiction.

D'après le tableau 2.1, nous avons présenté des études comparatives sur différentes classifieurs qui prouvent les performances des foréts aléatoires par rapport aux autres algorithmes de classification. Plusieurs études hybrides ont montré que l'utilisation de la forêt aléatoire combiné à d'autre technique d'apprentissage machine sert à améliorer la performance de la classification et la détection des problèmes dans le domaine médical. Nous avons présenté une vue d'ensemble de la forêt aléatoire et de sa performance dans la classification au domaine médical. Les forêts aléatoires sont rapides à construire et plus rapides à prévoir, et elles sont souvent plus précis qu'un seul classifieur.

5 Hybridation: Raisonnement à partir de cas et Forêts aléatoires

L'hybridation est une tendance visible dans les systèmes RàPC, bien qu'il existe de nombreux systèmes réussis basées sur l'approche RàPC standard. Ses performances peuvent être améliorées lorsqu'elles sont combinées par d'autres techniques d'apprentissage automatique ou d'exploration de données. Les classifications d'ensemble sont connues pour fournir une plus grande précision qu'un classifieur unique. Cependant, nous a motivé d'intégrer les forêts aléatoires comme mesure de similarité dans le système RàPC, pour but de modéliser la phase de remémoration et améliorer la qualité de la prise de décision des médecins. Très peu de travaux qui combinent le raisonnement à partir de cas avec les forêts aléatoires dans différents domaines, nous citons :

(Darabi et al, 2014), ont proposé le système RàPC pour le diagnostic de l'asthme, ils sont intégré les forêts aléatoires comme technique de sélection des caractéristiques pertinentes, les résultats ont montré que les caractéristiques sélectionnées par le système sont conformes au point devue d'un expert.

(Ayeldeen et al, 2015) ont combiné entre le système de raisonnement à partir de cas et les forêts aléatoires pour la prédiction du cancer du sein, ils sont utilisées les forêts aléatoires comme méthode d'optimisation, chaque cas dans la bibliothèque de cas est bien représenté et indexé pour la sélection des caractéristiques, le modèle proposé produit un haut degré d'efficacité avec une précision élevée.

(Zhong et al, 2015) ont développé un modèle hybride qui est basé sur les forêts aléatoires pour organiser les cas dans la base de cas. Le système est appliqué à la conception de générateur hydraulique, les résultats confirment que le modèle proposé est efficace pour la stabilité de la réutilisation des cas.

(Asim et al, 2019) ont développé un modèle prédictif et adaptatif appelé IB-RÀPC (Influential Blogger - Case Based Resonning) pour la reconnaissance des blogueurs influents invisibles. L'intégration des forêts aléatoires contribue à l'efficacité du modèle. Le modèle proposé présente un taux de précision de 88 à 95%.

(Tarchoune et al, 2021) ont combiné le RàPC et les algorithmes d'arbres de décision (C4.5, RepTree, LMT) et les forêts aléatoires classiques pour objectif de comparer la performance des algorithmes et de modéliser la phase de remémoration du système RàPC. Les résultats de simulation confirment la performance d'hybridation du système RàPC et les forêts aléatoires. Ils sont testé les algorithmes sur quatre bases de données médicales. Les résultats étaient satisfaisants.

(Tarchoune et al, 2022) ont intégré les forêts aléatoires modifiés dans la phase de remémoration du système RàPC, ils sont utilisés l'algorithme de forêt aléatoire de trois façon différentes : forêt aléatoire classique(CRF), forêt aléatoire avec sélection des attributs les plus importants (RF-FS) et forêt aléatoire pondérée (WRF), ils ont testé les trois algorithmes sur 11 bases de données médicales, les résultats montrent l'efficacité des algorithmes proposé pour modéliser la phase de remémoration.

(Tarchoune et al, 2023) ont présenté une revue systématique de raisonnement à partir de cas et les forêts aléatoires dans le domaine médical. Ils ont montré la performance de chaque méthode ainsi que la performance de l'hybridation de ces dernières.

6 Conclusion

Dans ce chapitre, nous avons présenté l'aspect théorique des différentes méthodes d'ensembles. Le choix des forêts aléatoires se justifie par les résultats obtenus par divers chercheurs dans la littérature scientifique. L'hybridation du raisonnement à partir de cas et les forêts aléatoires est une motivation pour intégrer ces dernières comme mesure de similarité dans un système de raisonnement à partir de cas (RàPC), afin de modéliser la phase de remémoration et d'améliorer la qualité des décisions. Ce sujet est détaillé dans le chapitre 4.

Chapitre 3

Techniques de sélection des caractéristiques

Chapitre 03

Techniques de sélection des caractéristiques

1 Introduction

La réduction de la dimensionnalité est devenue essentielle dans le domaine médical en raison de l'augmentation des données disponibles. Dans les systèmes d'aide au diagnostic médical, le système de résolution d'un problème repose souvent sur un ensemble de caractéristiques. Cependant, l'utilisation d'un grand nombre de caractéristiques peut nuire à la performance des algorithmes d'apprentissage en augmentant la complexité, le temps de calcul et les besoins en mémoire. L'objectif de la réduction de la dimensionnalité est de trouver une représentation plus compacte des données initiales dans un espace de dimensions réduites. Les techniques de réduction des caractéristiques se classifient généralement en deux catégories :

- Sélection de caractéristiques (Feature Selection) : Cette méthode consiste à choisir les caractéristiques pertinentes et non redondantes à partir de l'ensemble de données, en conservant uniquement celles qui décrivent le phénomène étudié de manière significative.

- Extraction de caractéristiques (Feature Extraction) : Cette approche remplace l'ensemble initial des données par un ensemble réduit, construit à partir des données initiales, en transformant les caractéristiques pour en créer de nouvelles qui capturent les informations essentielles.

Ce chapitre est structuré en trois sections : la première présente la définition des sélections des caractéristiques, la deuxième détaille la procédure de sélection des caractéristiques, et la troisième décrit les différentes méthodes de sélection de caractéristiques utilisées dans notre thèse.

2 Définition de la sélection des caractéristiques

La sélection des caractéristiques est le processus qui vise à identifier les caractéristiques pertinentes tout en éliminant celles qui sont non pertinentes, redondantes ou bruyantes. La littérature propose plusieurs définitions pour cette procédure.

La sélection des caractéristiques est décrite comme suit : à partir d'un ensemble de caractéristiques de dimension N, le processus de sélection vise à choisir un sous-ensemble qui minimise le taux d'erreur de classification.

(Dash et Liu, 1997) définissent le processus de sélection des caractéristiques en quatre étapes clés (voir Figure 3.1).

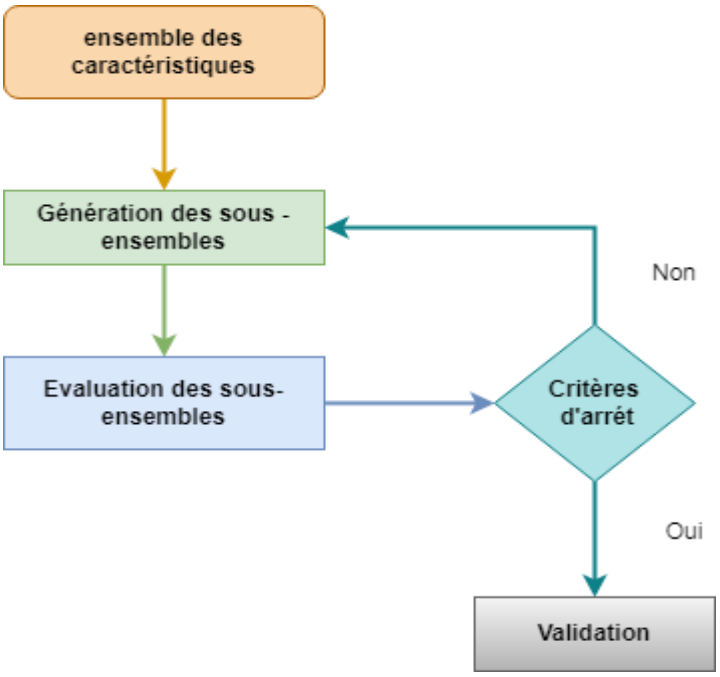


Figure 3.1. Représentation graphique du processus de sélection (Dash et Liu, 1997)

3 Procédure de la sélection des caractéristiques

Une procédure typique de sélection des caractéristiques, comme illustrée à la figure 3.1, se compose de quatre étapes : (i) génération des sous-ensembles, (ii) évaluation des sous-ensembles, (iii) application du critère d'arrêt, et (iv) validation des résultats.

1. Génération des sous-ensembles : Cette étape utilise une stratégie de recherche spécifique pour produire des sous-ensembles de caractéristiques candidats.

2. Évaluation des sous-ensembles : Chaque sous-ensemble candidat est évalué selon un critère d'évaluation défini, et les résultats sont comparés aux sous-ensembles précédemment examinés.

3. Critère d'arrêt : Le processus de génération et d'évaluation des sous-ensembles se poursuit jusqu'à ce qu'un critère d'arrêt prédéterminé soit atteint.

4. Validation des résultats : Le meilleur sous-ensemble de caractéristiques sélectionné est ensuite validé en utilisant des connaissances antérieures ou par des données de test.

Les stratégies de recherche et les critères d'évaluation sont des aspects cruciaux de la sélection des caractéristiques et influencent fortement l'efficacité du processus.

3.1 Génération

"À partir de quel point dans l'espace des caractéristiques la recherche doit-elle débuter ?"

Le choix du point de départ dans l'espace des sous-ensembles de caractéristiques peut influencer la direction de la recherche. Une fois le point de départ sélectionné de manière appropriée, une procédure de génération, également connue sous le nom de procédure de recherche, doit être définie. Trois principales stratégies de recherche sont couramment utilisées pour trouver un sous-ensemble optimal de caractéristiques :

1. Recherche exhaustive : Explore toutes les combinaisons possibles de sous-ensembles de caractéristiques pour identifier celui qui est optimal.
2. Recherche heuristique : Utilise des méthodes basées sur des règles empiriques ou des heuristiques pour guider la recherche de manière plus efficace, sans garantir une solution optimale mais en cherchant des solutions satisfaisantes.
3. Recherche aléatoire : Effectue une sélection aléatoire de sous-ensembles de caractéristiques pour découvrir des combinaisons potentiellement intéressantes.

3.2 Evaluation

L'évaluation de la performance des techniques de sélection de caractéristiques peut être abordée sous plusieurs angles. Dans cette sous-section, nous nous concentrerons sur les approches les plus couramment utilisées. En général, dans les problèmes de classification, notre objectif est de déterminer le sous-ensemble de caractéristiques le plus compact offrant la meilleure précision de classification. Les méthodes d'évaluation des caractéristiques se classifient principalement en trois catégories : les méthodes de filtrage (filter), les méthodes enveloppantes (wrapper), et les méthodes intégrées (embedded).

3.2.1 Méthode Filter

La méthode de filtrage, également connue sous le nom de modèle "filter", évalue la pertinence des caractéristiques en utilisant des critères basés sur les propriétés globales des données d'apprentissage, sans nécessiter l'intervention d'un algorithme d'apprentissage. Cette approche commence par choisir une méthode de recherche et définir la direction pour identifier les caractéristiques les plus pertinentes dans l'ensemble de données. Chaque caractéristique se voit ensuite attribuer un score de pertinence basé sur des mesures statistiques ; un score plus élevé indique une plus grande pertinence de la caractéristique (Saeys et al, 2007). Un aperçu visuel de cette méthode est Illustré dans la figure 3.2.

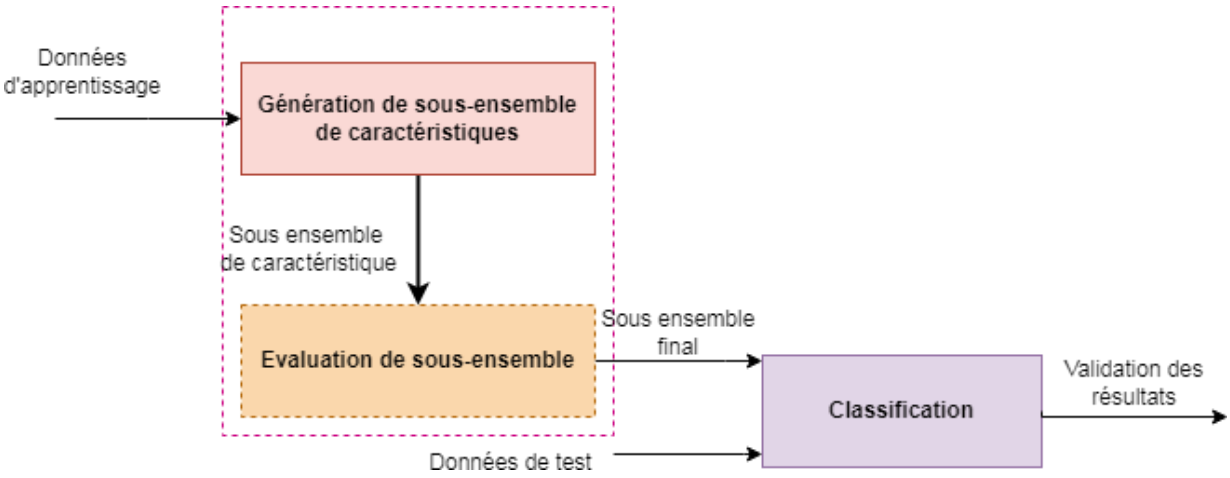


Figure 3.2. Illustration de modèle Filter (Tan, 2007)

Les méthodes de filtrage (filter) évaluent les caractéristiques en fonction de leur pertinence. Les caractéristiques obtenant les meilleurs scores sont considérées comme les plus pertinentes, tandis que celles avec des scores plus bas sont jugées moins importantes (Huang et al, 2007). En fin de processus, seules les caractéristiques ayant obtenu des scores élevés sont sélectionnés et utilisées comme entrées pour le classifieur (Saeys et al, 2007). Toutefois, une limitation significative des approches de filtrage est qu'elles ne tiennent pas compte de l'impact du sous-ensemble de caractéristiques sélectionnées sur la performance globale de l'algorithme d'apprentissage.

3.2.2 Méthodes Wrapper

Contrairement aux méthodes de filtrage qui déterminent un sous-ensemble optimal de caractéristiques sans tenir compte de l'algorithme d'apprentissage, les méthodes wrapper (ou enveloppantes) intègrent directement un algorithme d'apprentissage dans le processus de recherche du meilleur sous-ensemble de caractéristiques. Ces méthodes ont été largement adoptées en raison de leur capacité à offrir de meilleures performances en termes de généralisation (Bouaguel, 2015 ; Hewahi et Alashqa, 2015). Elles évaluent la qualité d'un sous-ensemble de caractéristiques en fonction des performances d'un classifieur spécifique (Kohavi et John, 1997). Une procédure de recherche est définie pour explorer l'espace des sous-ensembles possibles, différents sous-ensembles sont générés, et la précision de classification estimée par l'algorithme d'apprentissage est mesurée pour chaque sous-ensemble. Un aperçu des méthodes enveloppantes est illustré à la figure 3.3.

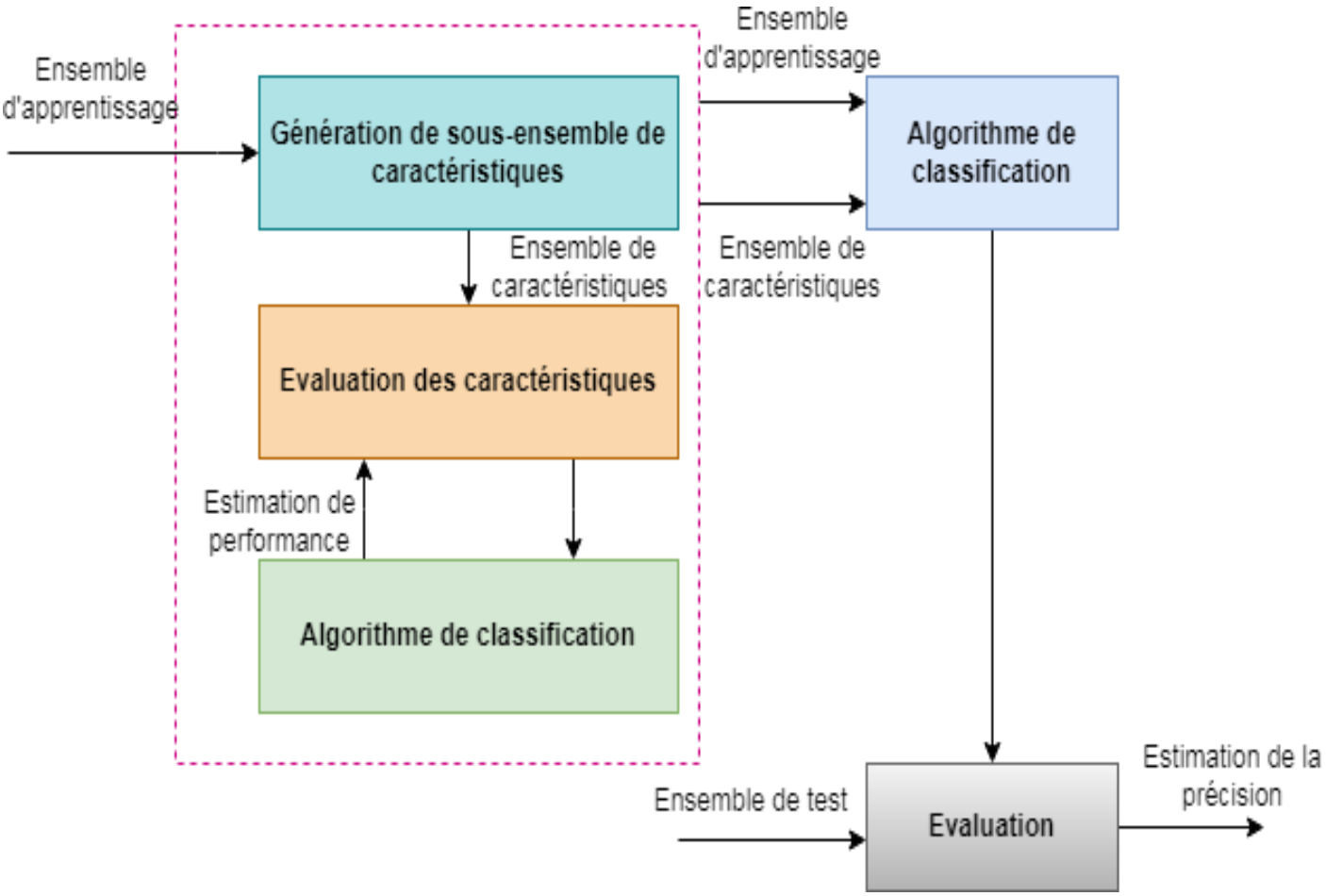


Figure 3.3. Illustration du modèle wrapper (Kohaviet John, 1997)

Les méthodes wrapper, en intégrant l'algorithme de classification dans le processus de sélection, permettent souvent d'atteindre une meilleure précision de classification des sous-ensembles choisis par rapport à ce que l'on obtient avec les méthodes de filtrage (Kohavi et John, 1997 ; Saeys et al, 2007). Cependant, ces sous-ensembles sont étroitement liés au classifieur utilisé, ce qui peut entraîner un risque accru de sur-apprentissage (overfitting) (Saeys et al, 2007). Certaines recherches (Stracuzzi, 2007) indiquent également que les approches wrapper non déterministes tendent à être plus rapides que leurs homologues déterministes.

3.2.3 Méthodes Embedded

Les méthodes intégrées, ou méthodes embarquées, combinent la sélection des caractéristiques directement dans le processus d'apprentissage du modèle. Contrairement aux méthodes wrapper, où les données d'apprentissage sont divisées en deux ensembles distincts — un pour l'apprentissage et un autre pour la validation du sous-ensemble de caractéristiques — les méthodes intégrées utilisent l'ensemble complet des données d'apprentissage pour optimiser le modèle. Cette approche permet (i) d'améliorer la qualité des résultats obtenus et (ii) de réduire le temps de calcul. Un schéma illustrant le processus de sélection des caractéristiques dans les méthodes intégrées est présenté dans la figure 3.4.

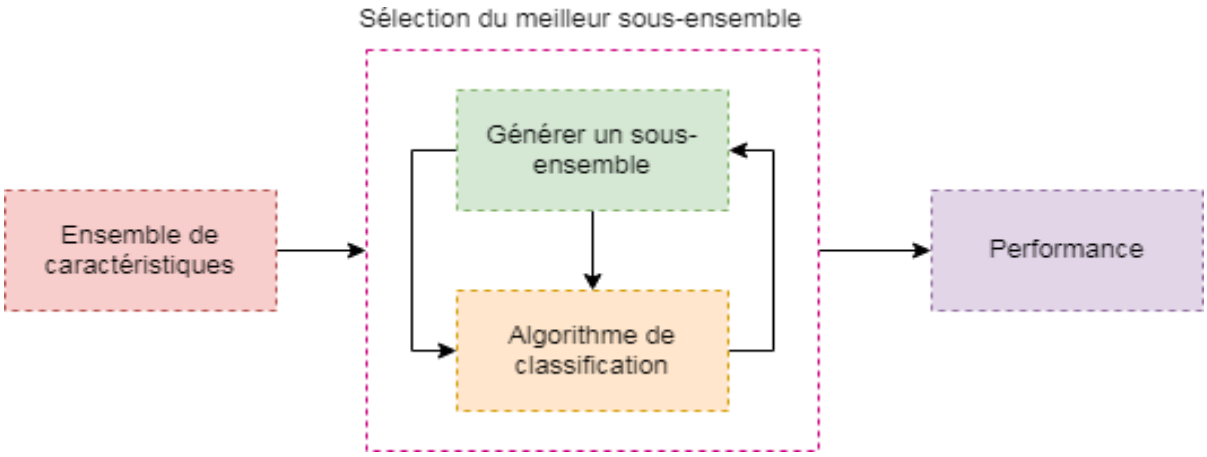


Figure 3.4. Illustration du modèle Embedded (Kaushik, 2016)

3.3 Critère d'arrêt

Le processus de sélection des caractéristiques doit être arrêté selon certains critères préétablis. Les critères pour mettre fin à la recherche du sous-ensemble optimal de caractéristiques peuvent inclure :

- La fin du processus de recherche.
- L'atteinte d'une taille préalablement définie pour le sous-ensemble de caractéristiques.
- L'exécution d'un nombre préétabli d'itérations.
- L'obtention d'un sous-ensemble de caractéristiques jugé optimal ou suffisamment bon selon les critères d'évaluation.
- L'absence d'amélioration significative de la qualité du sous-ensemble en cas d'ajout ou de suppression de caractéristiques.

3.4 Validation

Un moyen simple de valider les résultats est de comparer directement les résultats obtenus avec les connaissances préalables sur les caractéristiques des données. Si nous avons une connaissance préalable des caractéristiques pertinentes, nous pouvons comparer cet ensemble connu avec les caractéristiques sélectionnées. Les connaissances sur les caractéristiques non pertinentes ou redondantes peuvent également être utiles, mais il est peu probable qu'elles soient sélectionnées.

Dans les applications réelles, ces connaissances préalables ne sont généralement pas disponibles. Par conséquent, nous devons appuyer sur des méthodes indirectes pour évaluer la performance.Par exemple, nous pouvons surveiller le changement de performance en classification en comparant les taux d'erreur avant et après la sélection des caractéristiques. Pour cela, nous effectuons des expériences pour évaluer l'erreur de classification sur l'ensemble complet des caractéristiques et sur le sous-ensemble sélectionné.

4 Revue de quelques méthodes de sélection

Dans cette section, nous examinerons diverses méthodes de sélection des caractéristiques tirées de la littérature. Nous mettrons en avant des méthodes basées sur les différentes techniques de recherche mentionnées précédemment, ainsi que sur les diverses techniques d'évaluation.

4.1 Sélection par Corrélation

La sélection par corrélation est une méthode de filtrage simple qui évalue les sous-ensembles de caractéristiques en fonction de leur corrélation avec la classe cible, utilisant une fonction d'évaluation heuristique (Hall et al, 1999). Cette méthode privilégie les sous-ensembles de caractéristiques qui montrent une forte corrélation avec la classe tout en étant faiblement corrélées entre elles. Les caractéristiques non pertinentes, présentant une faible corrélation avec la classe cible, sont ainsi écartées (voir algorithme 3.1).

La fonction d'évaluation du sous-ensemble de fonctionnalités est calculée par la formule:

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (3.1)$$

Où Ms est le « mérite » heuristique d'un sous-ensemble de caractéristiques S contenant k caractéristiques, $\overline{r_{cf}}$ est la corrélation moyenne caractéristiques-classes (f∈S), et $\overline{r_{ff}}$ est l'intercorrélation moyenne caractéristiques-caractéristiques.

Algorithme 3.1: Sélection par Corrélation

Entrée : Ensemble de données d'apprentissage D, nombre de caractéristiques NF, seuil de corrélation.

Sortie : Caractéristiques sélectionnées

1. Déterminez l'ensemble des colonnes à inclure.
2. Calculez la matrice de corrélation entre les colonnes.
3. Parcourez la matrice de corrélation.
4. À chaque fois qu'une colonne a une corrélation supérieure au seuil avec l'une des colonnes déjà présentes, ajoutez-la à l'ensemble des colonnes sélectionnées.

4.2 Selection par Chi-square

La sélection par chi carré est décrite en détail par Zhu et al, 2007 et Bhalaji et al, 2018. Cette méthode est couramment utilisée pour tester l'indépendance entre deux événements. Les valeurs du chi carré sont calculées à l'aide de la formule suivante (voir algorithme 3.2).

$$x2 = \frac{(\mathrm{Oi}-\mathrm{Ei})\ 2}{\mathrm{Ei}} \qquad (3.2)$$

Où Oi est l'occurrence de l'attribut et Ei est l'occurrence de la classe.

Algorithme 3.2: Selection par Chi-square

Entrée : Ensemble de données d'apprentissage D, nombre de caractéristiques NF, valeurs x, valeurs y.

Sortie : Caractéristiques sélectionnées

1. Créez l'objet chi carré après l'entraînement avec les valeurs de x et y.
2. Calculez les valeurs du chi carré en utilisant la formule correspondante.
3. Triez les valeurs du chi carré de manière décroissante.
4. Supprimez la première colonne correspondant à cette valeur.
5. Répétez les étapes jusqu'à obtenir les caractéristiques pertinentes.

4.3 Sélection par Dropping Constant Features

Dans cette technique, la fonction de seuil de variance sert de sélecteur de caractéristiques en éliminant celles dont la variance est trop faible et qui n'apportent pas une valeur significative à la modélisation. Les caractéristiques dont la variance est proche de zéro sont supprimées. La variance de chaque caractéristique est calculée à l'aide de la formule suivante (voir algorithme 3.3).

$$\mathrm{Var(x)} = \frac{1}{n}\sum_{i=1}^{n}(xi - \bar{\mathrm{x}})^2 \qquad (3.3)$$

Ou xi est l'observation numéro i de la variable x et $\bar{\mathrm{x}}$ est la moyenne des xi

Algorithme 3.3: Sélection par Dropping Constant Features

Entrée: Jeu_de_données_d'entraînement_D, N_caractéristiques, seuil.

Sortie: Caractéristiques sélectionnées.

1. Créez un objet VarianceThreshold pour détecter les colonnes avec une faible variance selon le seuil.
2. Calculez les variances de chaque colonne.
3. Si cette colonne n'est pas dans la liste des colonnes à faible variance, ajoutez-la à la liste.
4. Supprimez la liste des colonnes à faible variance.

5 Discussions

Plusieurs techniques de sélection des caractéristiques ont été proposées dans la littérature pour réduire la taille des bases de données, L'utilisation de beaucoup de caractéristiques va influencer négativement sur la performance de l'algorithme d'apprentissage utilisé. La réduction de l'espace des caractéristiques est le processus consistant à identifier et supprimer autant d'informations redondantes que possible. Cela réduit la dimensionnalité des descripteurs et permet aux algorithmes d'apprentissage de fonctionner plus rapidement et plus efficacement.

La tâche d'un algorithme de sélection de caractéristiques est de fournir une solution computationnelle au problème de sélection de caractéristiques motivé par une certaine définition de la pertinence. Cet algorithme devrait être fiable et efficace. Les nombreux algorithmes de sélection des caractéristiques proposés dans la littérature reposent sur des principes très différents (la mesure d'évaluation utilisée, la manière précise d'explorer l'espace de recherche…etc.) et suivent de façon approximative différentes définitions de la pertinence.

6 Conclusion

La sélection des caractéristiques en apprentissage automatique est un domaine de recherche en pleine expansion, avec une multitude d'algorithmes variés en fonction de la procédure de recherche employée. Un facteur clé pour assurer la fiabilité des systèmes d'aide au diagnostic médical est la qualité et la représentation des données médicales. L'utilisation excessive de caractéristiques peut en effet nuire à la performance des algorithmes d'apprentissage.

Dans ce chapitre, nous avons défini la sélection des caractéristiques et exploré les méthodes les plus couramment utilisées : les méthodes filter, wrapper et embedded. La dernière section du chapitre propose un état de l'art sur les méthodes de sélection des caractéristiques, dont certaines seront approfondies dans les chapitres suivants.

Le chapitre suivant, fournit une vue plus détaillée des contributions réalisées ainsi que ses résultats obtenus.

Chapitre 4

Contributions

Chapitre 04

Contributions

4.1 Introduction

Le présent chapitre de recherche a pour objectif de concevoir une nouvelle approche hybride basée sur le raisonnement à partir de cas et les forêts aléatoires modifiées. Nous nous intéressons à la phase de remémoration ainsi qu'à la phase d'adaptation pour améliorer les performances du système RàPC. Dans cette section nous présentons les principales contributions.

4.2 Contribution1: Modélisation de la remémoration d'un RàPC par une forêt aléatoire améliorée et trois techniques de sélection des caractéristiques

Le raisonnement à partir de cas (RàPC) est fréquemment employé dans les systèmes d'aide à la décision médicale en raison de sa similitude avec le raisonnement humain. Bien que ce dernier soit efficace, ses performances peuvent être améliorées lorsqu'il est renforcé par d'autres techniques d'apprentissage automatique.

À cette fin, nous avons proposé un modèle hybride qui combine le RàPC et les forêts aléatoires. Tout d'abord, une étape de sélection des caractéristiques est réalisée en utilisant trois techniques : Corrélation, Dropping Constant Features et CHI2, afin de sélectionner les caractéristiques pertinentes. Ensuite, nous avons modélisé la phase de remémoration du système RàPC en utilisant des forêts aléatoires modifiées. Dans cette étape, deux méthodes d'élagage (pré-élagage et post-élagage) sont appliquées uniquement aux caractéristiques les plus importantes pour améliorer les performances du système RàPC.

Enfin, l'adaptation des cas remémorés est effectuée en fonction de règles d'adaptation proposées par un expert du domaine. L'approche développée a été évaluée sur 13 bases de données médicales, et les résultats ont montré que la méthode d'élagage proposée sélectionne le meilleur arbre en moins de temps et avec une meilleure précision par rapport à l'hybridation avec les forêts aléatoires classiques. Cela se traduit par une amélioration des performances du système RàPC.

4.2.1 Approche hybride combinant le raisonnement à partir de cas et les foréts aléatoires (CBR-RF)

Afin de concevoir une approche efficace pour réduire le temps et améliorer la précision des données médicales, nous proposons un modèle hybride combinant le raisonnement à partir de cas (RàPC) et les forêts aléatoires. Cette nouvelle méthode se distingue par l'amélioration des forêts aléatoires grâce à deux techniques d'élagage, le pré-élagage et le post-élagage, détaillées dans les sections 3.1 et 3.2 du chapitre 2.

L'organigramme illustré dans la figure 4.2.1 montre l'intégration des forêts aléatoires dans la première phase du système RàPC. Cette phase débute par des techniques de sélection visant à identifier les caractéristiques importantes et à éliminer celles qui sont redondantes ou non pertinentes. La sélection des caractéristiques est réalisée à l'aide de trois méthodes : Dropping Constant Features, Corrélation et Chi-square, décrites en détail dans les sections 4.1, 4.2 et 4.3 du chapitre 3.

Ensuite, nous avons intégré les forêts aléatoires améliorées ainsi que les forêts aléatoires classiques dans la phase de remémoration du système RàPC. Une fois cette phase achevée, nous avons procédé à l'adaptation des cas, où le cas remémoré est vérifié et comparé à la nouvelle règle correspondant au cas actuel.

La performance de l'approche proposée a été évaluée en utilisant les critères suivants : précision, rappel, score F1, et temps d'exécution sur 13 bases de données médicales.

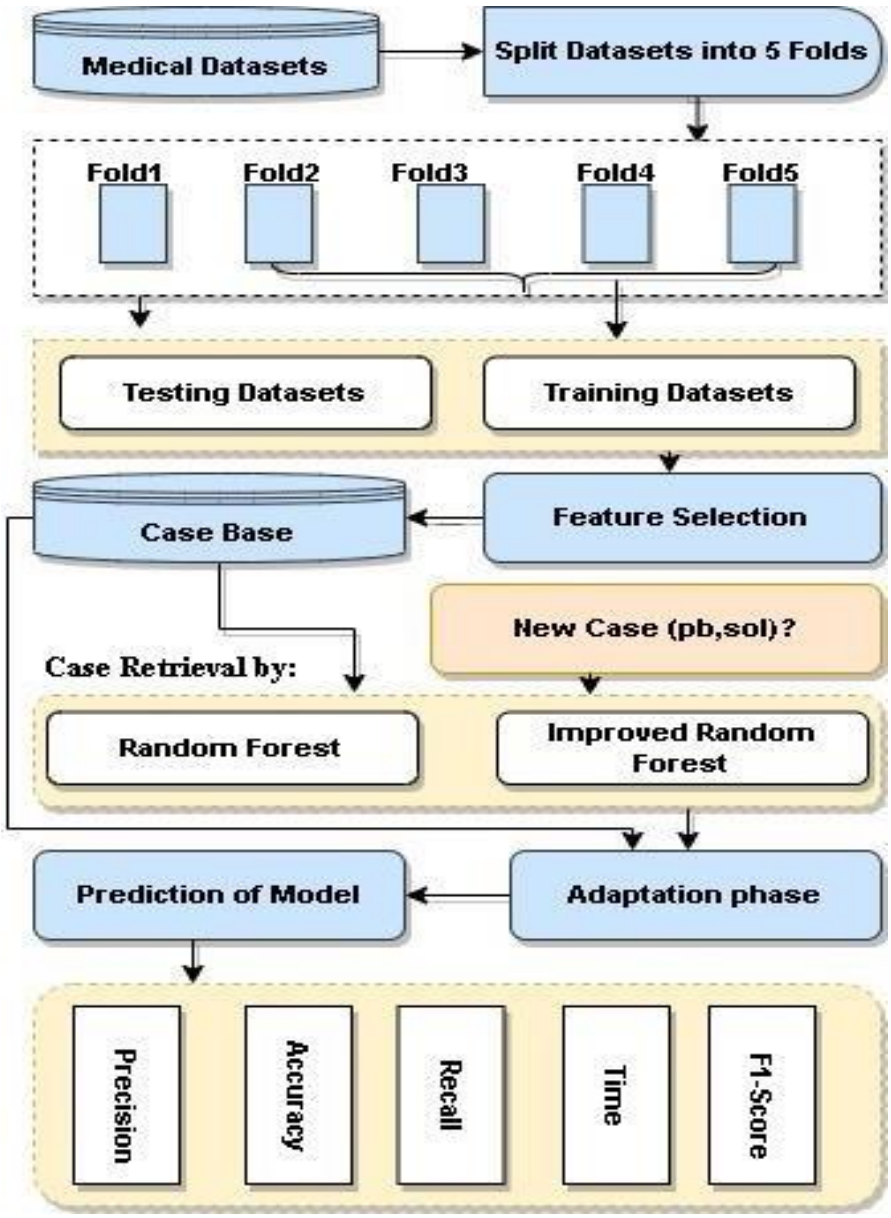


Figure 4.2.1. Organigramme de l'approche CBR-RF proposée (Tarchoune et al, 2024a)

4.2.2 Ensembles de données

Les ensembles de données utilisés dans cette expérience ont été collectés à partir du dépôt d'apprentissage automatique de l'Université de Californie à Irvine (UCI) et de Kaggle. Nous avons utilisé 13 bases de données médicales à deux classes. Le nombre d'échantillons dans ces ensembles de données varie de 100 à 1498, et le nombre de caractéristiques varie de 4 à 44.

Les détails spécifiques de ces ensembles de données sont présentés dans le tableau 4.2.1.

Tableau 4.2.1. Détails de l'ensemble de données et la taille de l'échantillon (Tarchoune et al, 2024a)

i	Base de données	# Les attributs	#les instances
1	Breast_Cancer_Wisconsin	9	699
2	Bupa	6	345
3	Diabetes	8	768
4	EEG-Eye-State	15	1498
5	haberman	4	306
6	Hepatitis	20	155
7	Parkinsons_LPD	22	195
8	Planning_Relax	13	303
9	Prostate_cancer	10	100
10	SaHeart	9	462
11	SPECTF_Heart	44	267
12	Statlog_Heart	14	303
13	Transfusion	5	748

L'ensemble de données a été divisé en ensembles d'entraînement et de test. Ensuite, nous avons utilisé l'ensemble d'entraînement avec une méthode de validation croisée en 5 fois pour optimiser le modèle.

4.2.3 Étape principales du CBR-RF proposées

L'algorithme proposé de notre approche est détaillé dans cette section (Figure 4.2.2), il exécute les étapes suivantes:

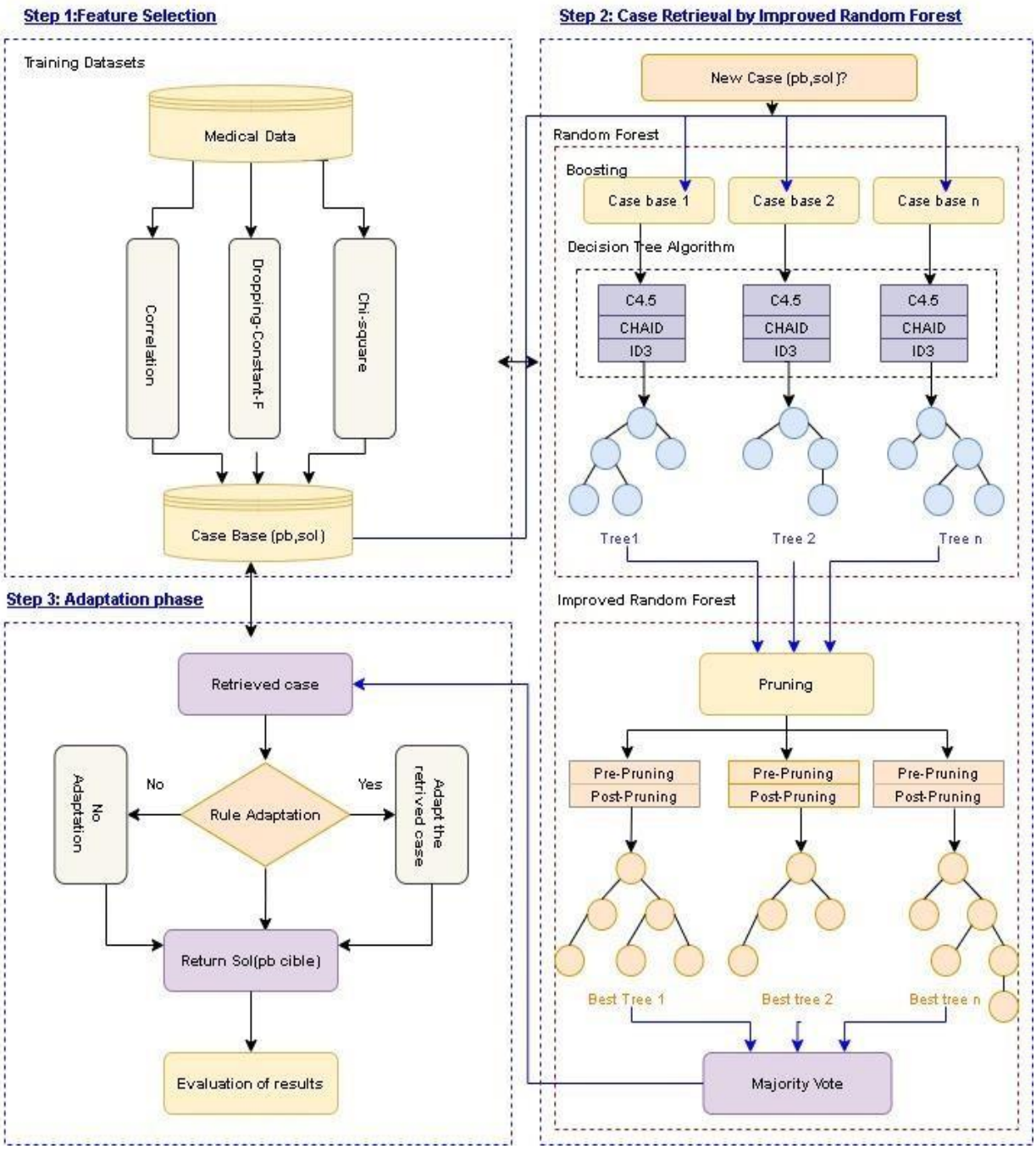


Figure 4.2.2. L'architecture détaillée de l'approche proposée (Tarchoune et al, 2024a)

Étape 1: Sélection des caractéristiques

Dans de nombreux scénarios d'apprentissage, l'obtention d'un ensemble minimal de caractéristiques avec une bonne capacité de prédiction est parfois plus avantageuse que le simple développement de modèles. La sélection des caractéristiques permet d'identifier celles ayant les scores les plus élevés tout en éliminant les caractéristiques redondantes et non pertinentes. Cela améliore l'aspect pratique pour les cliniciens et augmente la performance de prédiction (Djellali et al, 2018; Speiser, 2021; Williamson et al, 2022; Zemmal et al, 2019; Zhou et al, 2021). Les algorithmes de sélection des caractéristiques sont souvent indispensables pour le développement de la base de cas. Dans notre étude, nous avons appliqué trois techniques de sélection des caractéristiques pour chaque maladie : Dropping Constant Features, Corrélation et Chi2-square.

L'objectif de cette section est de réduire la dimensionnalité et de sélectionner les caractéristiques pertinentes pour chaque maladie. Une fois les méthodes de sélection des caractéristiques terminées, la base de cas du système RàPC est construite.

Étape 2: Remémoration par des forêts aléatoires

Modélisation de la phase de remémoration par l'intégration des forêts aléatoires. Tout d'abord, nous avons appliqué les forêts aléatoires classiques en utilisant trois algorithmes de construction des arbres de décision (ID3, C4.5, CHAID) sur les caractéristiques hautement classées. Ensuite, nous avons amélioré la phase de remémoration en utilisant des forêts aléatoires modifiées basées sur deux méthodes d'élagage (pré-élagage et post-élagage). Enfin, nous avons comparé le comportement de la remémoration en utilisant les forêts aléatoires classiques avec celui des forêts aléatoires améliorées.

A. Forêt Aléatoire classique

Dans cette section, nous avons utilisé l'algorithme de forêt aléatoire classique basé sur l'arbre de décision. Dans la littérature, différents algorithmes d'arbre de décision ont été appliqués, notamment ID3, C4.5 et CHAID. Nous proposons d'utiliser ces algorithmes sur plusieurs bases de données pour sélectionner le meilleur arbre comme classifieur individuel dans la forêt. Ensuite, nous évaluons la performance des trois algorithmes afin de choisir le plus approprié, dans le but d'augmenter la performance de classification des forêts aléatoires.

1. Algorithme ID3

L'algorithme ID3 a été développé par Ross Quinlan. Dans cet algorithme, l'attribut avec le gain d'information le plus élevé est considéré comme le plus informatif et est sélectionné comme nœud racine de l'arbre (Zhang et al., 2021). Ce processus est répété jusqu'à ce que tous les attributs soient inclus dans l'arbre. La construction de l'arbre de décision se déroule comme suit : tout d'abord, l'entropie de la classe et l'entropie de chaque attribut sont calculées. Ensuite, le gain d'information est déterminé pour tous les attributs, comme le décrit l'équation 4.2.1 et 4.2.2 ci-dessous (Hssina et al, 2014).

$$Entropie(p) = -\sum_{i=1}^{n} pi * \log(pi) \qquad (4.2.1)$$

Où valeurs (pi) est l'ensemble de toutes les valeurs possibles pour un attribut T.

$$Gain(p,T) = \text{Entropie(p)} - \sum_{j=1}^{n}(\text{pj} * \text{Entropie(pj)}) \qquad (4.2.2)$$

Où :

- Gain (p, T) représente le gain d'information pour un test T et une position p,
- Entropie(p) est l'entropie de la classe p,
- pj est la proportion d'instances pour chaque valeur possible vj d'un attribut T,
- Entropie (pj) est l'entropie de la classe pour chaque valeur vj.

Cette mesure est utilisée pour classer les attributs et construire l'arbre de décision en sélectionnant ceux qui maximisent le gain d'information.

1. Algorithme C4.5

L'algorithme C4.5 a été développé pour surmonter certaines des limitations de l'algorithme ID3. L'une des principales améliorations apportées par C4.5 est l'utilisation du **gain ratio**, un critère de sélection moins biaisé (Hssina et al, 2014). Le gain ratio prend en compte le gain d'information et le normalise en fonction de l'entropie fractionnée.

En C4.5, deux nouveaux paramètres sont calculés en plus de ceux de l'équation 4.2.2,

à savoir :

$$GainRatio(p,T) = \frac{\text{Gain(p, T)}}{Splitinfo\text{(p, T)}} \qquad (4.2.3)$$

Where Split Info is:

$$Splitinfo\text{(p, T)} = -\sum_{J=1}^{n} p'\left(\frac{j}{p}\right) * \log\left(\text{p}'\left(\frac{j}{p}\right)\right) \qquad (4.2.4)$$

P’ (j/p) est la proportion d’éléments présents à la position p, en prenant la valeur du j-ième test.

2. Algorithme CHAID

CHAID est un algorithme de construction d'arbre développé par kass1980 (Kass, 1980)**,** s'appuie sur le test du khi-deux par la formule (Girard, 2007):

$$x^2 = \sum_{k=1}^{k} \sum_{l=1}^{L} \frac{nkl - \frac{nk*nl}{n}}{\frac{nk*nl}{n}} \qquad (4.2.5)$$

Où

- nkl représente le nombre d'éléments appartenant à la classe Yk pour la valeur xl de l'attribut.
- nk est la somme du nombre d'éléments de toutes les classes pour une certaine valeur de l'attribut.
- nl est la somme du nombre d'éléments pour toutes les valeurs possibles de l'attribut, calculée pour une certaine classe.
- n est le nombre total d'exemples dans le jeu de données d'apprentissage.

Ces variables sont utilisées pour calculer l'entropie et le gain d'information, permettant ainsi d'évaluer la pertinence des attributs dans la construction de l'arbre de décision.

B. Forêt Aléatoire améliorée proposée

Malgré l'efficacité démontrée des forêts aléatoires dans diverses applications (Anirudh et al, 2019; Divya et al, 2019; Proniewska et al, 2020; Yu et al, 2020; Alam et al, 2020; Li et al, 2020; Raposo et al, 2020; Asadi et al, 2021), plusieurs chercheurs ont tenté d'améliorer la précision en utilisant uniquement les meilleurs arbres de la forêt. Cependant, pour optimiser notre modèle de forêt aléatoire, nous avons proposé deux méthodes d'élagage afin de réduire la taille du modèle et le risque de sur-ajustement. L'objectif de cette section est de mettre en évidence différentes techniques d'élagage et de comparer leur performance.

L'algorithme 4.2.1 présente notre approche modifiée pour la phase de remémoration en utilisant les forêts aléatoires améliorées.

Algorithme 4.2.1: Remémoration basée sur les forêts aléatoires améliorées (Proposées) (Tarchoune et al, 2024a)

Entrée: M échantillons et N caractéristiques.

Sortie : Forêt_aléatoire_améliorée

1. Rééchantillonnez l'échantillon M fois pour obtenir des sous-jeux de données aléatoires pour l'entraînement.

2. Construisez un arbre (ID3, C4.5 et CHAID) pour chaque sous-ensemble de données.

3. Répétez n fois pour obtenir n arbres de décision.

4.Élaguez les arbres de décision selon deux types: élagage préliminaire et élagage postérieur.

5. Supprimez les sous-arbres superflus.

6. Obtenez les meilleurs arbres de décision pour chaque échantillon.

7. Combinez les arbres dans une nouvelle forêt aléatoire.

Les algorithmes de forêts aléatoires étendent les concepts des arbres de décision pour construire des modèles prédictifs plus stables. Pour construire chacun de ces arbres, des échantillons sont tirés avec remplacement à partir des données. Par conséquent, certains échantillons peuvent être utilisés plusieurs fois dans le même arbre. Le problème de sur-apprentissage est souvent rencontré lors de la construction des arbres. En effet, un arbre de décision peut surajuster les données d'entraînement s'il est autorisé à atteindre sa profondeur maximale. La restriction de la profondeur maximale est une méthode simple pour simplifier l'arbre et gérer le sur-ajustement. L'objectif principal de cette section est d'aborder ce défi à l'aide de deux méthodes d'élagage, qui consistent à supprimer les sous arbres superflus, afin d'optimiser le temps de calcul et d'améliorer les performances.

Plusieurs méthodes d'élagage sont décrites dans la littérature (Esposito et al, 1993; Helmbold et Schapire, 1995; Mohamed et al, 2012; Nan et al, 2016; Wardhani et al, 2022). Ces méthodes diffèrent par leur stratégie d'élagage et le type d'élagage utilisé. Dans cette étape, nous présentons deux types d'élagage basés sur les règles générées à partir d'un arbre de décision : le pré-élagage (élagage au moment de la construction de l'arbre) et le post-élagage (élagage après la construction de l'arbre).

Étape 3: Phase d'Adaptation

Une fois le cas remémoré trouvé, nous l'adaptons en utilisant des règles traitées par un expert. Enfin, nous avons évalué les résultats obtenus sur 13 bases de données médicales. Étant donné que l'adaptation est particulièrement complexe en médecine, nous avons proposé une méthode de remémoration guidée par l'adaptation automatique des cas existants en fonction des règles modifiées automatiquement par l'expert du domaine. Dans notre approche, lorsqu'un cas remémoré est trouvé, nous recherchons tous les symptômes correspondants à ce cas et comparons le cas remémoré avec le cas cible.

L'algorithme 4.2.2 décrit notre approche pour soutenir la phase d'adaptation :

1. Identification des caractéristiques importantes : Tout d'abord, les caractéristiques importantes de chaque maladie sont précisées par un expert.
2. Application des règles d'adaptation : Si le cas remémoré ne satisfait pas les contraintes du cas cible, nous appliquons les règles d'adaptation pour ajuster la solution du cas source afin qu'elle corresponde au cas cible. Si la solution est déjà satisfaisante, aucune adaptation n'est nécessaire.

Algorithme 4.2.2: Phase d'adaptation proposée (Tarchoune et al, 2024a)

Entrée : Cas récupéré, L1 : Liste des caractéristiques importantes, L2 : Liste des caractéristiques du cas récupéré.

Sortie : Cas adapté.

Algorithme :

1. Pour i de 1 à N faire ; Pour j de 1 à M faire ;
2. Initialiser :
   - Nb.count ← 0 // Nombre de caractéristiques importantes.
   - L2.count ← nombre de caractéristiques du cas récupéré.
3. Pour i de 1 à L2.count faire :
   - Si le caractéristique dans L1 alors
     - Nb.count ← Nb.count + 1.
4. Si Nb.count ≤ L2.count alors
   - Aucune adaptation nécessaire. Sinon
   - Adapter (cas récupéré) // Remplacer la classe du cas récupéré par la classe réelle.

4.2.4 Expérimentations et évaluations

Cette section décrit l'organisation générale des expériences, pour démontrer la performance de la méthode proposée, nous avons implémenté nos algorithmes sur 13 bases de données médicales. La première série d'expériences a été réalisée pour valider les techniques de sélection des caractéristiques, dans ce qui suit, impact des forêts aléatoires amélioré sur la phase deremémoration du système RàPC. Enfin la proposition d'un algorithme de modélisation de la phase d'adaptation.

4.2.4.1 Performance de la sélection des caractéristiques

Dans cette section, nous avons évalué trois méthodes différentes de sélection des caractéristiques et trois algorithmes de construction des arbres pour les forêts aléatoires. Le tableau 4.2.2 présente la précision des algorithmes de forêt aléatoire classique (utilisant C4.5, CHAID, ID3) dans deux conditions : avec et sans utilisation des techniques de sélection des caractéristiques, sur 13 bases de données médicales. Chaque type d'algorithme d'arbre de décision utilise un processus de construction différent et donne des résultats variés. En général, la majorité des bases de données (13 bases de

données) ont montré une amélioration de la précision des forêts aléatoires classiques grâce à l'utilisation des techniques de sélection des caractéristiques.

Tout d'abord, la mise en œuvre de la technique du chi carré a généralement donné les meilleures précisions sur la plupart des bases de données, indiquant que la méthode chi carré est compatible avec toutes les bases de données et les types d'algorithmes utilisés. Ensuite, la technique "Dropping Constant Features" s'est révélée être la meilleure dans 6 bases de données lorsqu'elle était utilisée avec l'algorithme C4.5 pour construire les arbres de la forêt, avec une précision maximale de 100%. Enfin, la technique de "Corrélation" a été la meilleure dans 3 bases de données lorsque l'algorithme C4.5 était appliqué.

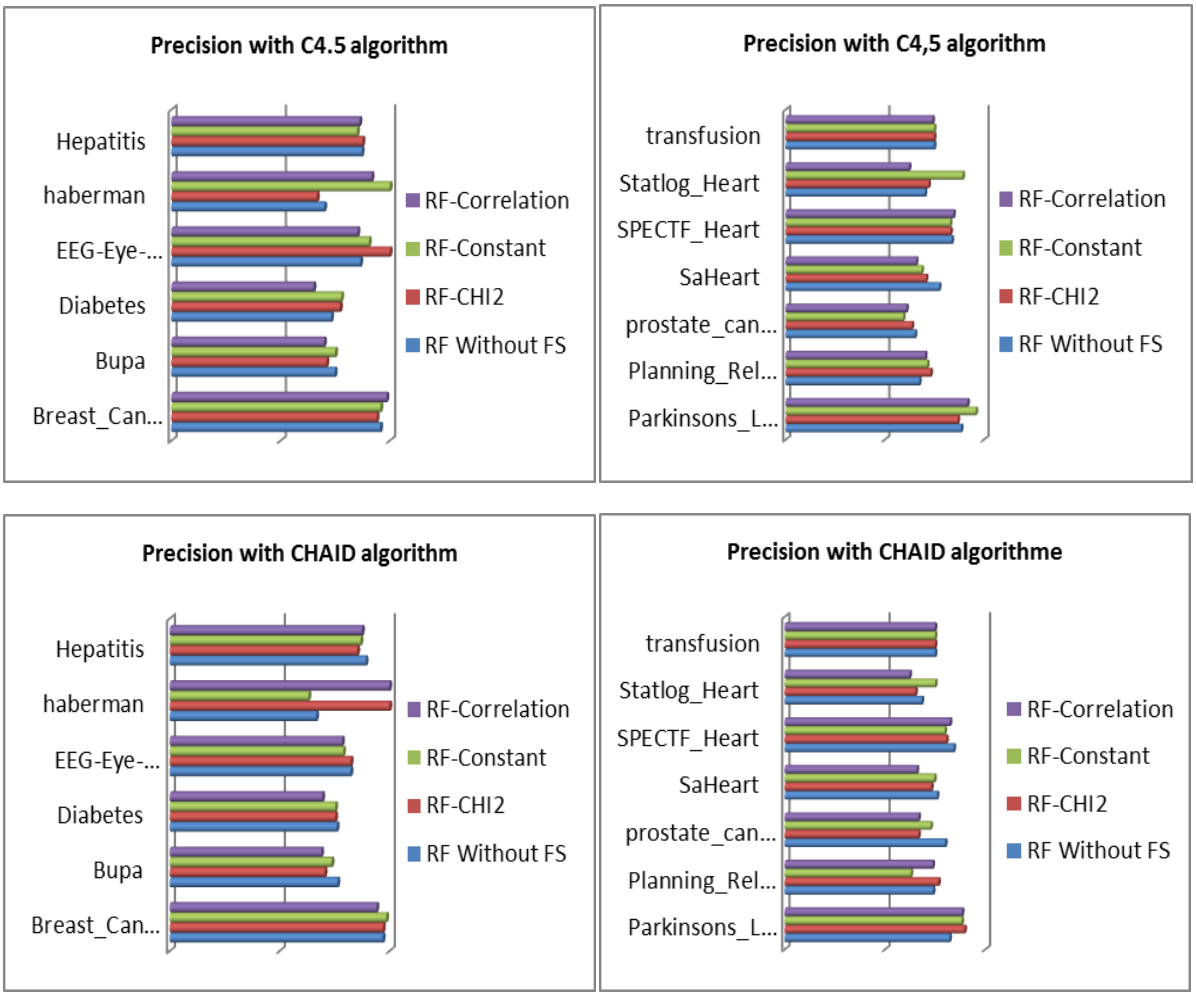
Tableau 4.2.2. Nombre de base de données ayant la meilleure ou la plus mauvaise précision de classification (Tarchoune et al, 2024a)

Methodes / Nombre de base/ Precision	CBR-RF			CBR-RF-Chi2			CBR-RF-Constant			CBR-RF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Best datasets	0	1	0	**5**	**4**	**3**	**3**	**6**	**1**	**2**	**3**	**1**
Worst datasets	12	12	12	**1**	**1**	**1**	**2**	**2**	**2**	**5**	**5**	**5**

- Lorsque le nombre de caractéristiques est élevé, les performances de classification tendent d'abord à s'améliorer. Par exemple, les bases de données EEG-Eye-state, Haberman et Parkinson ont obtenu une précision de 100%.
- À mesure que la taille de la base de données et le nombre de caractéristiques augmentent, l'utilisation des techniques de sélection devient cruciale et est considérée comme une étape majeure dans le processus de classification.
- Les méthodes de sélection (Chi-square, Dropping Constant Feature, Correlation) sont efficaces selon l'algorithme utilisé et la base de données. Par exemple, la base de données Parkinson a obtenu la meilleure précision de 95.92% avec l'algorithme C4.5 et la technique Dropping Constant Feature.
- Quel que soit la technique utilisée, les résultats sont restés pratiquement performants avec l'algorithme correspondant.
- Lorsque le nombre de caractéristiques est très faible, c'est-à-dire que les caractéristiques sont pertinentes dès le départ, une certaine dégradation de la performance est observée, comme dans le cas de la base de données transfusion.

Tableau 4.2.3. Comparaison de la précision obtenue par de différentes méthodes avec et sans sélection des caractéristiques (Tarchoune et al, 2024a)

Base de données \ Methodes	CBR-RF			CBR-RF-Chi2			CBR-RF-Constant			CBR-RF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Breast_Cancer_Wisconsin	95,59	96,97	94,29	94.03	97.06	**98.33**	95.71	**98.53**	98.48	**98,51**	94.12	97.10
Bupa	74,80	76,34	72,59	71.01	70.50	**74.42**	75.00	73.68	**77.78**	69.93	68.97	69.93
Diabetes	73,17	76,03	77,21	**77.18**	75.32	76.76	**77.78**	75.32	77.18	65.10	69.43	**77.88**
EEG-Eye-State	86,49	82,30	82,28	**100.00**	82.36	76.81	**90.48**	78.93	83.20	85.19	78.41	74.90
Haberman	70,00	66,67	100,00	66.67	**100.00**	83.33	**100.00**	63.16	68.18	91.67	**100.00**	61.11
Hepatitis	87,18	89,19	88,89	87.50	85.37	88.57	85.00	86.84	**91.43**	86.11	87.50	**91.67**
Parkinsons_LPD	88,46	82,14	87,50	86.79	89.58	**92.45**	**95.92**	88.24	89.09	**91.67**	88.37	**93.62**
Planning_Relax	67,39	73,91	71,70	73.08	**76.60**	68.09	**71.43**	62.86	71.11	70.37	73.68	71.43
Prostate_cancer	65,22	80,00	69,57	63.64	66.67	69.57	59.26	72.73	66.67	60.87	66.67	66.67
SaHeart	77,39	75,89	71,54	70.77	73.04	77.57	68.57	74.58	**72.17**	65.82	65.82	70.23
SPECTF_Heart	83,75	84,21	82,56	83.13	80.56	**83.13**	82.93	79.73	81.40	**84.62**	82.35	81.18
Statlog_Heart	70,21	68,29	83,87	**72.09**	65.12	69.23	**89.29**	75.00	75.00	62.22	62.22	62.22
Transfusion	74.90	74.90	77.33	74.90	74.90	74.90	74.90	74.90	77.33	74.09	74.90	74.90



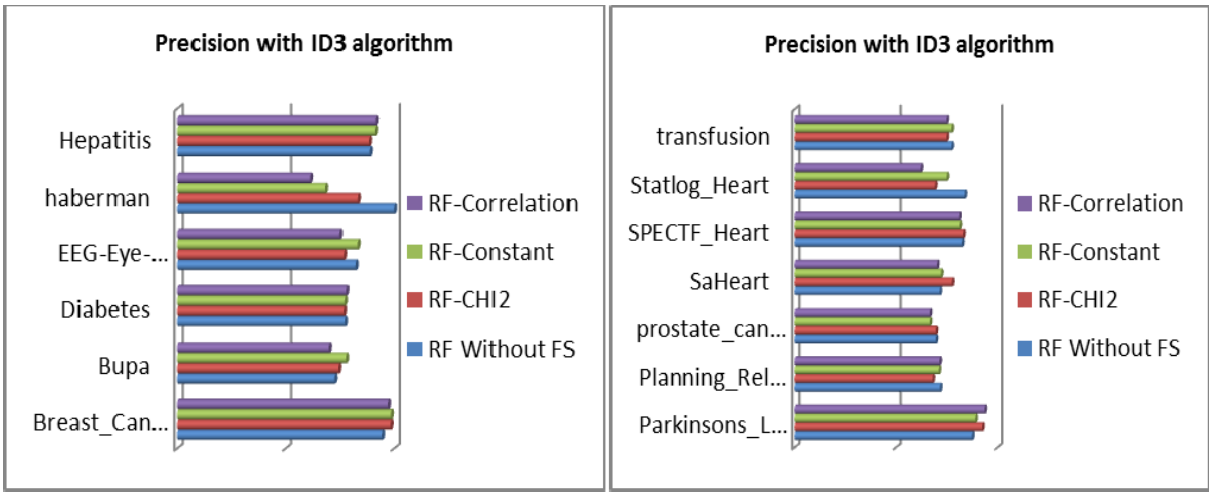


Figure 4.2.3. Performance des méthodes proposés avec et sans sélection des caractéristiques (Tarchoune et al, 2024a)

Nous avons observé que, en plus de fournir une précision élevée dans la majorité des algorithmes de classification, les techniques de sélection présentent également l'avantage de réduire le temps de calcul. Ainsi, les méthodes de sélection jouent un rôle crucial lorsqu'un ensemble de données comporte de nombreuses caractéristiques. En particulier, l'algorithme chi carré sélectionne un nombre minimal de caractéristiques tout en améliorant les performances, tandis que l'algorithme de Correlation est privilégié en raison de son temps d'exécution très court.

Tableau 4.2.4. Comparaison du temps obtenue à partir de différentes méthodes avec et sans sélection des caractéristiques (Tarchoune et al, 2024a)

Methodes / Base des données	CBR-RF			CBR-RF-Chi2			CBR-RF-Constant			CBR-RF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Breast_Cancer_Wisconsin	29,05	29,05	29,05	**22,84**	**20,44**	**22,90**	27,35	23,98	26,99	26,74	21,33	23,77
Bupa	25,26	25,26	25,26	19,98	14,77	20,01	26,83	21,61	26,91	**7,96**	**6,93**	**8,66**
Diabetes	60,28	60,28	60,28	48,77	48,14	79,92	53,26	54,51	78,03	**3,85**	**3,48**	**3,84**
EEG-Eye-State	145,97	145,97	145,97	129,84	150,10	164,20	144,30	165,30	162,10	**100,70**	**140,50**	**120,20**
Haberman	12,53	12,53	12,53	**8,17**	**6,36**	**7,85**	11,76	9,50	12,61	12,18	10,25	14,93
Hepatitis	6,57	6,57	6,57	9,87	7,39	8,70	6,61	4,65	5,46	**4,93**	**3,79**	**4,57**
Parkinsons_LPD	9,11	9,11	9,11	69,84	51,17	63,62	21,29	16,61	20,08	80,04	45,27	50,26
Planning_Relax	69,50	69,50	69,50	48,92	48,63	41,87	52,23	43,32	56,08	**34,44**	**27,18**	**30,42**
Prostate_cancer	51,43	51,43	51,43	14,80	11,01	12,36	8,07	8,30	10,02	**8,47**	**6,13**	**7,39**
SaHeart	16,66	16,66	16,66	40,92	44,57	54,58	41,34	37,71	61,68	**6,75**	**5,37**	**7,52**
SPECTF_Heart	50,39	50,39	50,39	45,23	149,29	129,32	46,10	185,78	156,85	50,10	152,03	124,25
Statlog_Heart	70,10	70,10	70,10	42,40	31,53	33,95	33,20	24,04	30,48	**3,18**	**3,21**	**3,14**
Transfusion	20,84	20,84	20,84	20,84	18,56	27,03	21,67	17,44	28,25	20,84	18,56	27,03

D'après le tableau 4.2.4, le temps de calcul est amélioré dans la plupart des bases de données, avec la technique de Correlation atteignant la meilleure performance sur 8 bases de données. Cette analyse nous permet de conclure que les méthodes de sélection des caractéristiques ont produit de meilleurs résultats par rapport aux forêts aléatoires classiques.

Le temps de calcul a été mesuré pour tous les scénarios, tant avec l'utilisation des techniques de sélection que sans. Le tableau 4.2.4 montre le temps de calcul (en secondes) des trois algorithmes avec les trois techniques de sélection proposées. Selon ce tableau, la technique de Correlation avec l'algorithme CHAID est la plus efficace (requiert moins de temps) en termes de temps de calcul sur 10 bases de données. Ces résultats sont prometteurs, car ils indiquent une amélioration du temps de calcul sans perte de précision. Cela suggère que la réduction du nombre de caractéristiques prises en compte dans le calcul de similarité réduit effectivement le temps nécessaire pour évaluer la similarité entre deux cas, optimisant ainsi la base de cas du système RàPC.

4.2.4.2 Rémémoration par des forêts aléatoires améliorées

Après l'élimination des caractéristiques non pertinentes, le modèle de la forêt aléatoire modifié a été exécuté et évalué à nouveau sur 13 bases de données pour valider l'efficacité de la méthode proposée.

4.2.4.2.1 Forêts aléatoires améliorées avec Pré-élagage

La comparaison des différentes méthodes en termes de précision est également présentée dans le tableau 4.2.5. Ce tableau montre que l'approche CBR-IRF surpasse systématiquement le système RàPC classique en termes de précision, indépendamment de la taille des données. En utilisant des méthodes plus adaptées pour traiter des types de données complexes, l'approche CBR-IRF améliore les performances par rapport aux méthodes traditionnelles de raisonnement à partir de cas, comme illustré dans la figure 4.2.4.

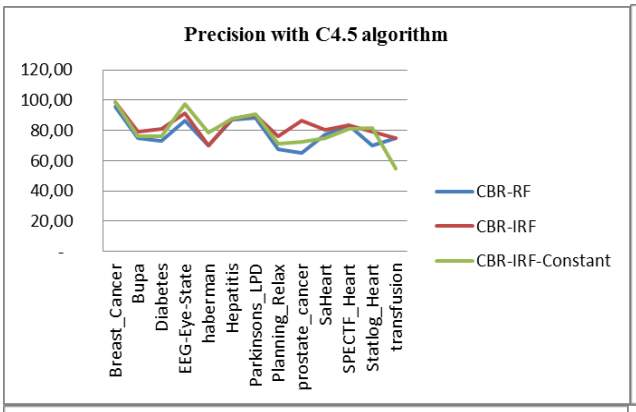
D'après le tableau 4.2.5, il est clair que parmi toutes les méthodes testées, l'approche CBR-IRF-CHI2 avec l'algorithme CHAID sur la base de données Breast-cancer-wisconsin a atteint la meilleure précision avec 98,52%. De même, l'approche CBR-IRF-Constant avec l'algorithme C4.5 a obtenu la meilleure précision de 97,77% sur la base EEG-Eye-State. Il est notable que l'approche proposée a constamment surpassé le CBR-RF avec différents algorithmes et tailles de données. Cela peut être expliqué comme suit :

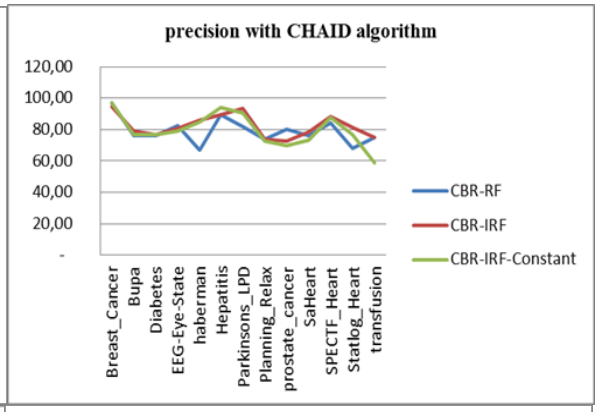
- **Pré-élagage efficace :** L'algorithme de pré-élagage proposé donne de meilleurs résultats pour la majorité des bases de données. En choisissant et en sélectionnant les meilleurs arbres, l'approche améliore les performances globales, le pré-élagage offrant ainsi le meilleur compromis en termes de performance.
- **Amélioration par les forêts aléatoires :** L'intégration des forêts aléatoires améliorées dans la phase de remémoration du système RàPC améliore encore la précision de prédiction. Cela démontre que l'approche CBR-IRF fonctionne bien avec de grands ensembles de données, offrant à la fois une précision élevée et une efficacité suffisante.

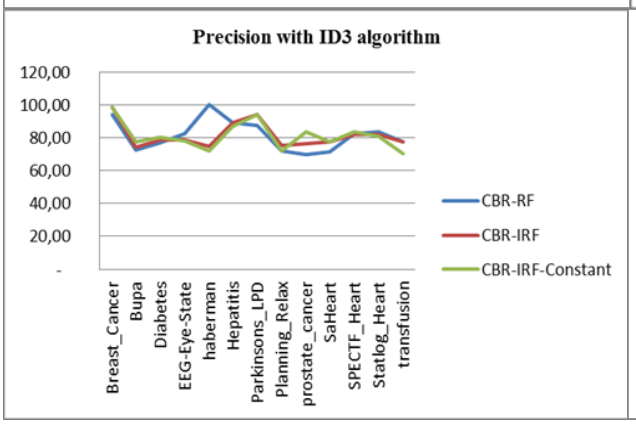
Tableau 4.2.5. Performance des méthodes d'apprentissage appliqué sur 13 bases de données médicales

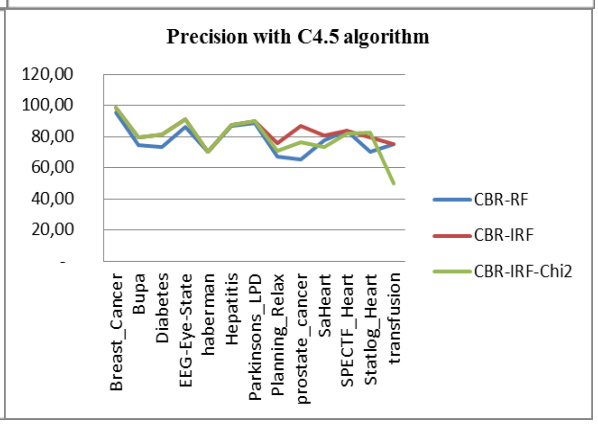
(Tarchoune et al, 2024a)

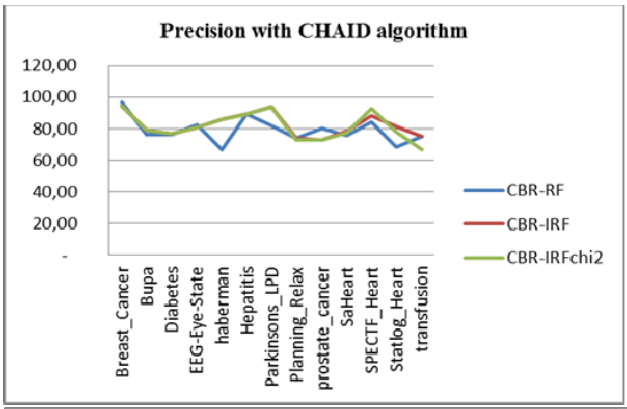
Methodes / Base de données	CBR-RF			CBR-IRF-Chi2			CBR-IRF-Constant			CBR-IRF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Breast_Cancer_Wisconsin	95,59	96,97	94,29	**98,51**	94,44	**98,51**	**98,51**	97,06	98,48	**98,48**	**98,53**	95,65
Bupa	74,80	76,34	72,59	**79,34**	**79,03**	74,05	75,97	76,74	77,52	72,99	69,93	74,07
Diabetes	73,17	76,03	77,21	**81,34**	76,82	78,36	76,28	76,92	**80,00**	65,10	77,88	69,43
EEG-Eye-State	86,49	82,30	82,28	**91,43**	81,06	78,60	**97,78**	79,09	78,28	**94,29**	71,70	76,63
Haberman	70,00	66,67	100,00	70,00	**85,71**	75,00	78,57	**84,62**	72,00	68,97	**77,78**	75,00
Hepatitis	87,18	89,19	88,89	87,50	89,47	89,19	87,50	**94,29**	87,18	87,50	87,50	87,50
Parkinsons_LPD	88,46	82,14	87,50	90,20	**93,75**	**94,00**	90,74	90,38	**94,00**	90,57	89,09	90,57
Planning_Relax	67,39	73,91	71,70	**71,19**	72,73	73,33	**71,19**	72,88	71,93	**71,70**	74,51	73,77
Prostate_cancer	65,22	80,00	69,57	76,19	72,73	**78,95**	72,73	69,57	**83,33**	72,73	76,19	66,67
SaHeart	77,39	75,89	71,54	73,48	77,12	**79,21**	75,00	73,13	**77,59**	66,24	74,81	67,81
SPECTF_Heart	83,75	84,21	82,56	82,14	**92,96**	90,28	81,18	**87,50**	83,72	88,31	85,19	87,34
Statlog_Heart	70,21	68,29	83,87	82,35	**78,05**	82,35	81,82	76,92	80,95	62,22	62,22	62,22
Transfusion	74,90	74,90	77,33	50,00	66,67	66,67	54,55	58,82	70,37	50,00	66,67	100,00

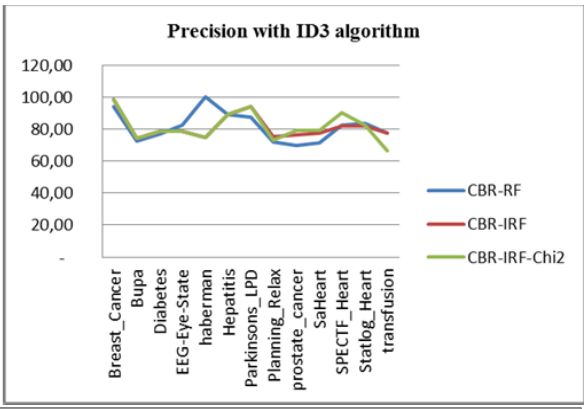

Precision with C4.5 algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-Constant

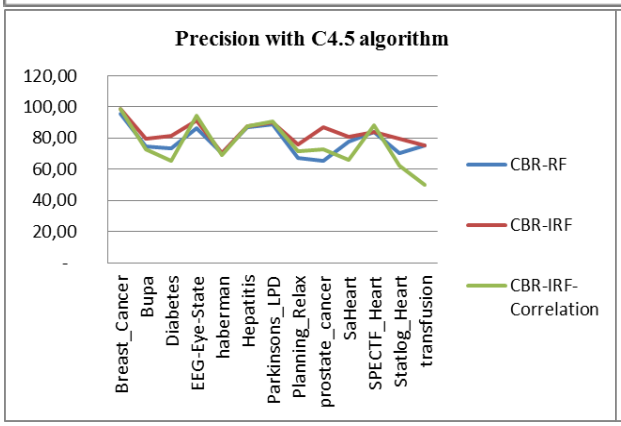

precision with CHAID algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-Constant

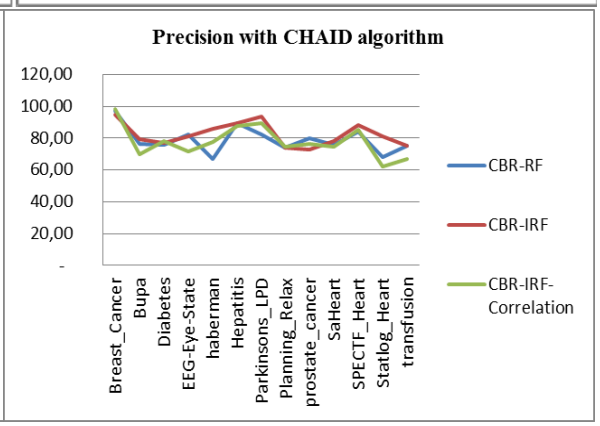

Precision with ID3 algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-Constant


Precision with C4.5 algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-Chi2


Precision with CHAID algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRFchi2


Precision with ID3 algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-Chi2


Precision with C4.5 algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-
Correlation


Precision with CHAID algorithm
120,00
100,00
80,00
60,00
40,00
20,00
-
Breast_Cancer
Bupa
Diabetes
EEG-Eye-State
haberman
Hepatitis
Parkinsons_LPD
Planning_Relax
prostate_cancer
SaHeart
SPECTF_Heart
Statlog_Heart
transfusion
CBR-RF
CBR-IRF
CBR-IRF-
Correlation

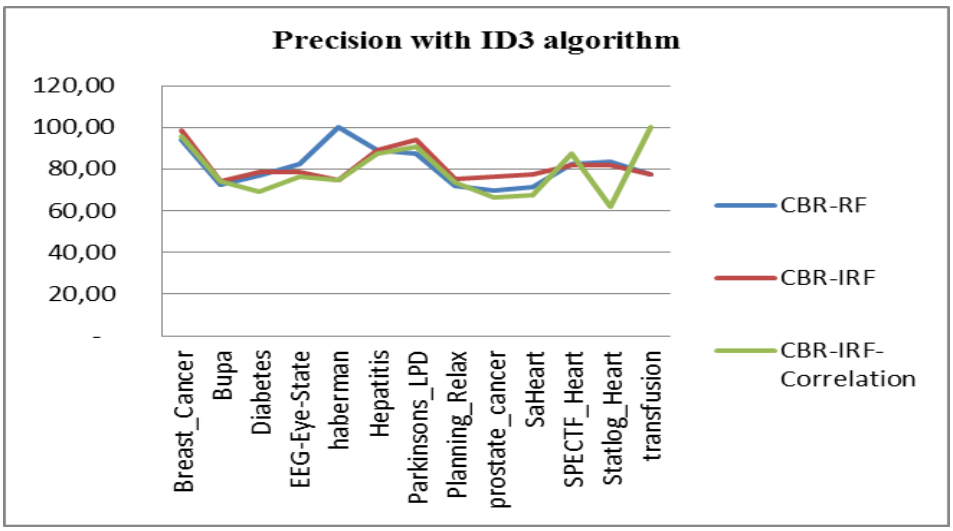


Figure 4.2.4. Comparaison de la précision obtenue par de différentes méthodes proposées sur 13 bases de données médicales (Tarchoune et al, 2024a)

Pour construire un modèle efficace avec un temps d'apprentissage réduit, nous avons opté pour l'application de l'algorithme de pré-élagage après avoir utilisé les techniques de sélection de caractéristiques. Le tableau 4.2.6 présente une comparaison des performances des différentes méthodes en termes de temps de calcul. La figure 4.2.6 montre le temps de calcul pour chaque méthode sur toutes les bases de données, avec les principales observations suivantes :

- **Variation du temps de calcul :** Le temps de calcul varie en fonction de la taille de la base de données, des caractéristiques sélectionnées et de l'algorithme appliqué.
- **Performance des techniques de sélection :** La technique de Correlation montre le temps d'exécution la plus faible sur 9 bases de données, tandis que la technique Chi-square est la moins performante sur 2 bases de données. En comparaison, la technique Dropping Constant Feature est la plus lente parmi les techniques proposées.

Cependant, l'évolution de la performance des forêts aléatoires en fonction des techniques de sélection et de pré-élagage est très efficace, tant en termes de précision que de temps de calcul. L'objectif de cette étape est de comparer et de mettre en évidence l'amélioration des forêts aléatoires classiques grâce à l'utilisation de caractéristiques pertinentes et à la sélection des meilleurs arbres.

Tableau 4.2.6. Comparaison du temps obtenue à partir de différentes méthodes avec et sans sélectiondes caractéristiques (Tarchoune et al, 2024a)

Methodes / Base de données	CBR-RF			CBR-IRF-Chi2			CBR-IRF-Constant			CBR-IRF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Breast_Cancer_Wisconsin	29,05	29,05	29,05	**27,89**	**22,79**	**23,54**	26,05	26,68	28,95	27,81	26,82	24,79
Bupa	25,26	25,26	25,26	16,35	15,87	17,80	20,39	20,36	23,68	**7,99**	**4,88**	**8,94**
Diabetes	60,28	60,28	60,28	36,41	57,05	65,21	49,09	43,79	63,12	**3,40**	**3,14**	**3,37**
EEG-Eye-State	145,97	145,97	145,97	103,37	133,40	129,40	110,26	130,50	127,10	**99,09**	**123,50**	**122,90**
haberman	12,53	12,53	12,53	**5,41**	**5,23**	**7,96**	9,19	10,56	13,48	9,80	8,33	10,51
Hepatitis	6,57	6,57	6,57	8,42	7,38	9,39	5,05	5,32	5,71	**5,01**	**4,01**	**5,05**
Parkinsons_LPD	9,11	9,11	9,11	49,34	54,85	49,75	11,42	13,76	13,76	50,61	62,34	32,88
Planning_Relax	69,50	69,50	69,50	57,70	35,19	40,83	43,56	36,85	35,82	**24,91**	**21,81**	**28,24**
prostate_cancer	51,43	51,43	51,43	13,65	11,33	12,84	7,63	8,42	11,19	**5,90**	**5,54**	**6,86**
SaHeart	16,66	16,66	16,66	29,44	72,79	59,86	47,06	32,25	60,59	**5,09**	**6,21**	**5,69**
SPECTF_Heart	50,39	50,39	50,39	218,28	143,07	145,22	161,94	150,01	143,65	179,37	120,82	113,40
Statlog_Heart	70,10	70,10	70,10	33,04	30,00	31,18	40,61	24,06	27,77	**2,75**	**2,81**	**2,75**
transfusion	20,84	20,84	20,84	20,84	18,23	26,93	20,24	17,56	26,83	**19,84**	**17,56**	**25,03**

4.2.4.2.2 Forêts aléatoires améliorées avec Post-élagage

Dans cette section, la comparaison des méthodes est présentée dans le tableau 4.2.7. Les résultats montrent que la méthode d'élagage (Post-élagage) donne également d'excellents résultats en termes de précision pour la plupart des bases de données. Voici les principales observations :

- **Performance générale :** La méthode CBR-IRF utilisant le Post-élagage offre une performance moyenne par rapport aux résultats obtenus avec CBR-RF. Cependant, certains résultats sont particulièrement notables :
  - Sur la base Breast Cancer, l'approche CBR-IRF-Constant avec l'algorithme CHAID atteint une précision de 98,53%.
  - Pour la base Parkinsons-LPD, l'approche CBR-IRF-Constant avec l'algorithme C4.5 obtient une précision de 95,92%.
- **Comparaison avec les méthodes classiques :** Dans la plupart des bases de données, la précision obtenue avec les méthodes d'élagage est presque similaire à celle des méthodes classiques utilisant la sélection des caractéristiques.
- **Efficacité du pré-élagage :** Il est important de noter que, pour les résultats mentionnés, la méthode Post-élagage est moins performante que le pré-élagage en termes de précision de classification. Le pré-élagage semble offrir un meilleur compromis en termes de précision globale.

Tableau 4.2.7. Performance des méthodes d'apprentissage appliqué sur 13 bases de données médicales (Tarchoune et al, 2024a)

Methodes / **Bases de données**	**CBR-RF**			**CBR-IRF-Chi2**			**CBR-IRF-Constant**			**CBR-IRF-Correlation**		
	C45	**CHAID**	**ID3**	**C45**	**CHAID**	**ID3**	**C45**	**CHAID**	**ID3**	**C45**	**CHAID**	**ID3**
Breast_Cancer_Wisconsin	95,59	96,97	94,29	94,03	97,06	98,33	**95,71**	**98,53**	**98,48**	98,51	94,12	97,10
Bupa	74,80	76,34	72,59	71,01	70,50	74,42	75,00	73,68	**77,78**	69,93	68,97	69,93
Diabetes	73,17	76,03	77,21	**77,18**	75,32	76,76	**77,78**	75,32	77,18	65,10	69,43	77,88
EEG-Eye-State	86,49	82,30	82,28	**100,00**	82,36	76,81	**90,48**	78,93	83,20	85,19	78,41	74,90
Haberman	70,00	66,67	100,00	66,67	100,00	83,33	100,00	63,16	68,18	**91,67**	**100,00**	61,11
Hepatitis	87,18	89,19	88,89	87,50	85,37	88,57	85,00	86,84	**91,43**	86,11	87,50	**91,67**
Parkinsons_LPD	88,46	82,14	87,50	86,79	**89,58**	**92,45**	**95,92**	**88,24**	**89,09**	91,67	88,37	93,62
Planning_Relax	67,39	73,91	71,70	73,08	**76,60**	68,09	71,43	62,86	71,11	70,37	73,68	71,43
Prostate_cancer	65,22	80,00	69,57	63,64	66,67	69,57	59,26	72,73	66,67	60,87	66,67	66,67
SaHeart	77,39	75,89	71,54	70,77	73,04	**77,57**	68,57	74,58	72,17	65,82	65,82	70,23
SPECTF_Heart	83,75	84,21	82,56	83,13	80,56	83,13	82,93	79,73	81,40	**84,62**	82,35	81,18
Statlog_Heart	70,21	68,29	83,87	72,09	65,12	69,23	**89,29**	**75,00**	75,00	62,22	62,22	62,22
Transfusion	74,90	74,90	77,33	74,90	74,90	74,90	74,90	74,90	77,33	74,09	74,90	74,90

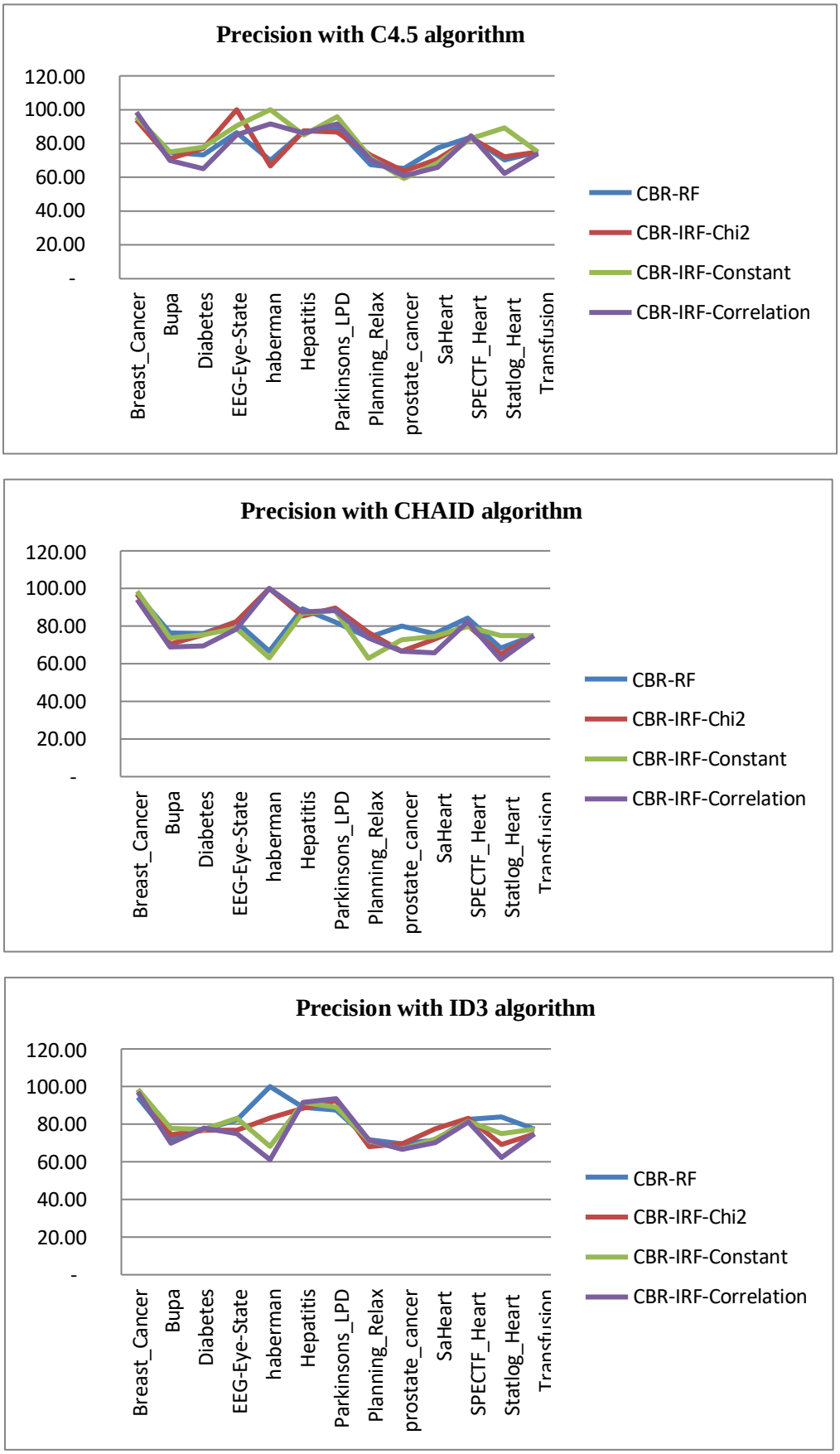
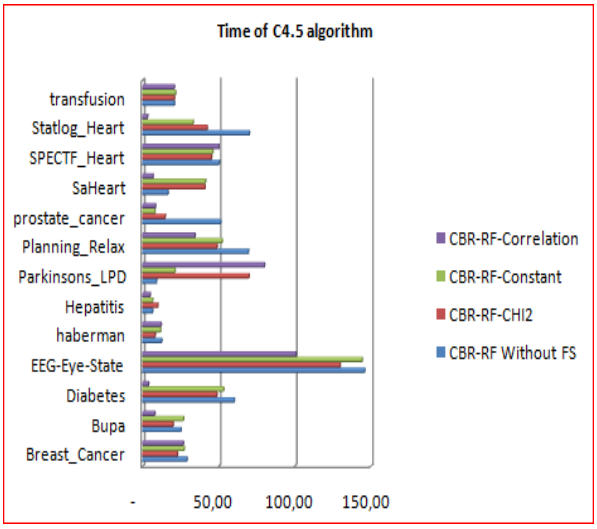
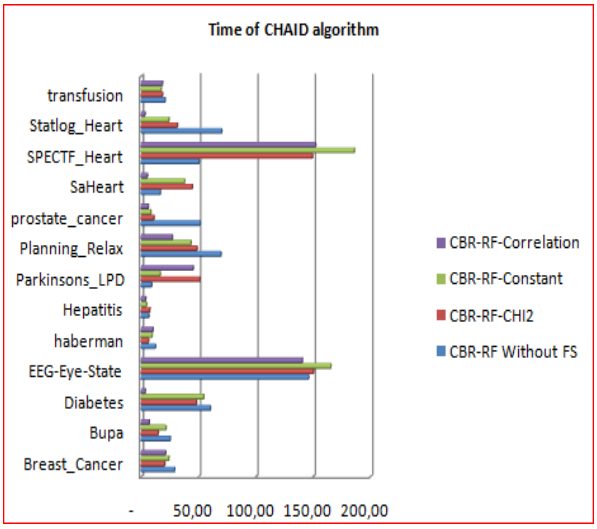
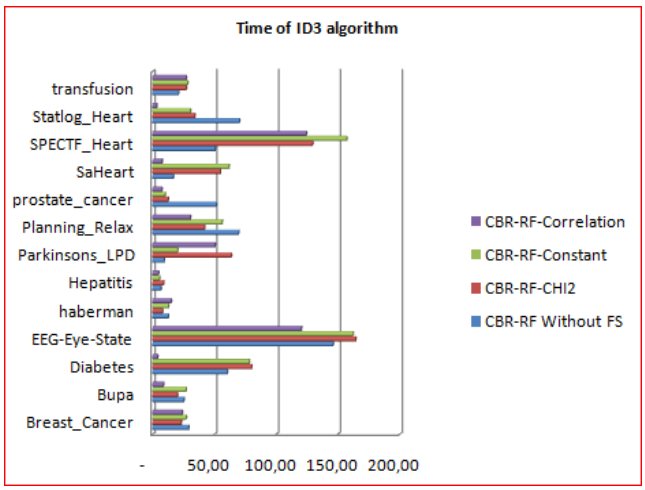


Figure 4.2.5. Performance des méthodes d'apprentissage appliqué sur 13 bases de données médical

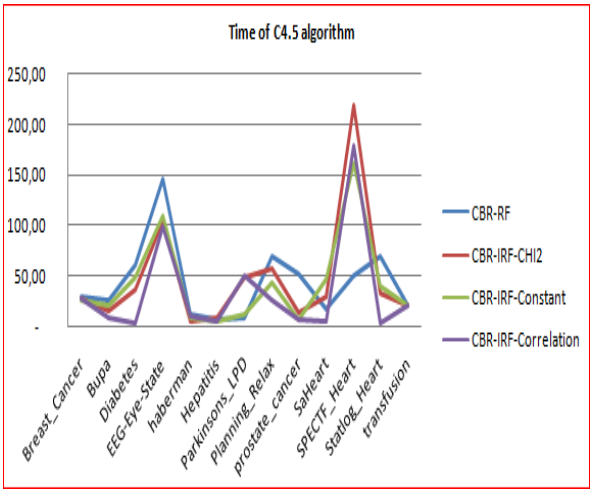
Intégration de la forêt aléatoire classique dans le raisonnement à partir de cas avec et sans sélection de caractéristiques

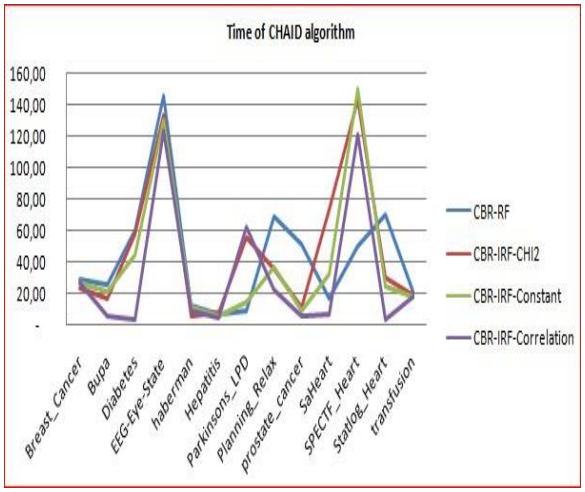


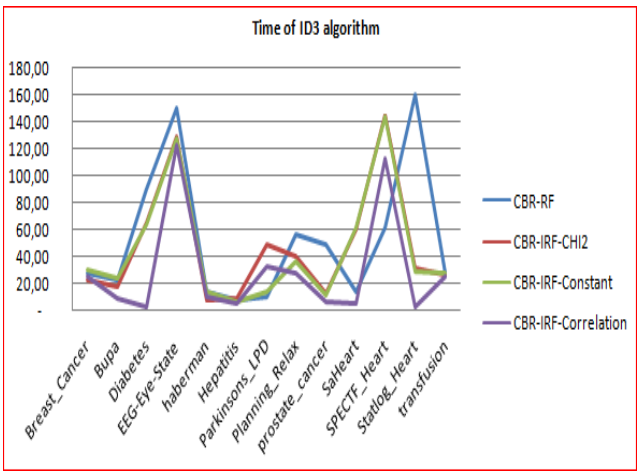




Intégration de la forêt aléatoire améliorée (pré-élagage) dans le raisonnement à partir de cas avec sélection de caractéristiques







Intégration de la forêt aléatoire améliorée (post-élagage) dans le raisonnement à partir de cas avec sélection de caractéristiques

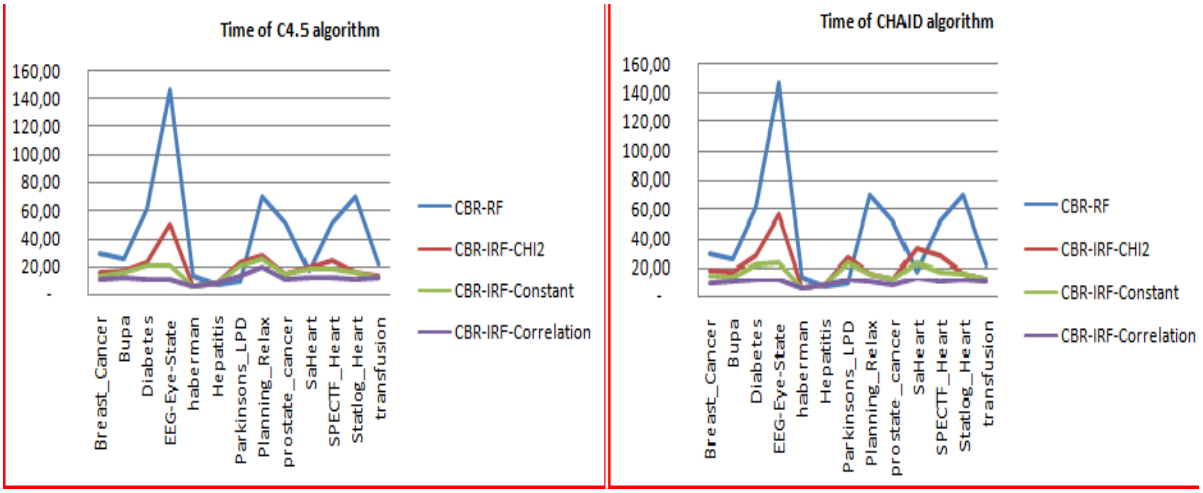


Figure 4.2.6. Évaluation du temps des méthodes d'apprentissage appliqué sur 13 bases de données médical (Tarchoune et al, 2024a)

D'après le tableau 4.2.8, les analyses comparatives des méthodes décrites dans cette section montrent que :

- **Performance de la sélection de caractéristiques :** L'approche de sélection de caractéristiques basée sur la Corrélation (CBR-IRF-Correlation) atteint la meilleure performance en termes de temps de calcul sur toutes les bases de données, quel que soit l'algorithme utilisé (C4.5, CHAID, ID3).
- **Comparaison des méthodes d'élagage :**
  - La méthode de Post-élagage est plus performante en termes de temps de calcul par rapport à la méthode de Pré-élagage.
  - En revanche, la méthode de Pré-élagage surpasse le Post-élagage en termes de précision.

En résumé, les deux méthodes d'élagage présentent des avantages distincts :

- **Pré-élagage** : Offre une meilleure précision.
- **Post-élagage** : Réduit le temps de calcul.

Les performances globales en termes de précision et de temps de calcul dépendent donc de l'équilibre entre la technique de sélection des caractéristiques, la méthode d'élagage, et l'algorithme appliqué.

Tableau 4.2.8. Comparaison du temps obtenue à partir de différentes méthodes avec et sans sélection des caractéristiques (Tarchoune et al, 2024a)

Methodes / Base de données	CBR-RF			CBR-IRF-Chi2			CBR-IRF-Constant			CBR-IRF-Correlation		
	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3	C45	CHAID	ID3
Breast_Cancer_Wisconsin	29,05	29,05	29,05	15.89	17.79	16.54	12.89	13.79	14.54	**9.89**	**8.79**	**10.54**
Bupa	25,26	25,26	25,26	16,35	15,87	17,80	15.35	12.87	14.80	**11.35**	**10.87**	**11.80**
Diabetes	60,28	60,28	60,28	22.41	28.05	23.21	20.41	22.05	19.21	**10.41**	**12.05**	**11.21**
EEG-Eye-State	145,9	145,97	145,97	50.37	56.40	43.40	20.37	23.40	19.40	**10.37**	**11.40**	**10.40**
Haberman	12,53	12,53	12,53	5,41	5,23	7,96	**5.41**	**5.23**	**7.96**	5.41	5.23	7.96
Hepatitis	6,57	6,57	6,57	8,42	7,38	9,39	8.42	7.38	9.39	**8.42**	**7.38**	**9.39**
Parkinsons_LPD	9,11	9,11	9,11	22.34	25.85	19.75	20.34	22.85	19.75	12.34	11.85	9.75
Planning_Relax	69,50	69,50	69,50	**27.70**	**15.19**	**21.83**	25.70	15.19	21.83	**18.70**	**10.19**	**12.83**
Prostate_cancer	51,43	51,43	51,43	13,65	11,33	12,84	14.65	12.33	12.84	**9.65**	**7.33**	**7.84**
SaHeart	16,66	16,66	16,66	19,44	32,79	18,86	18.44	22.79	12.86	**11.44**	**12.79**	**13.86**
SPECTF_Heart	50,39	50,39	50,39	**24.28**	**28.07**	**22.22**	18.28	17.07	12.22	11.28	10.07	7.22
Statlog_Heart	70,10	70,10	70,10	**15.04**	**15.00**	**12.18**	15.04	15.00	11.18	**10.04**	**12.00**	**9.18**
Transfusion	20,84	20,84	20,84	**12.84**	**11.23**	**13.93**	11.84	11.23	13.93	**11.84**	**10.23**	**9.93**

4.2.4.3 Évaluation de la phase d'adaptation

Dans cette section, nous comparons la performance de l'approche proposée dans deux conditions : avec et sans l'algorithme d'adaptation, sur les 13 bases de données médicales.

La phase d'adaptation, intégrée à notre approche, dépend du nombre de caractéristiques présentes dans chaque base de données. Pour identifier les cas remémorés adaptés, nous avons développé l'algorithme 7, qui procède comme suit :

1. Création des Listes : Les listes L1 (caractéristiques sélectionnées par l'expert) et L2 (caractéristiques du cas remémoré) sont créées.

2. Comparaison des Listes : Les deux listes sont comparées. Si les caractéristiques du cas remémoré satisfont les conditions définies par L1, aucune adaptation n'est nécessaire. Sinon, la classe du cas remémoré est remplacée par la classe réelle.

Exemple d'Application : Pour la base de données Diabetes, les quatre caractéristiques les plus importantes retournées par la phase de remémoration sont listées dans L2. Si L1 et L2 partagent trois caractéristiques, et si le nombre de caractéristiques de L1 est supérieur à celui de L2, nous adaptons le cas remémoré en remplaçant la classe par la classe réelle.

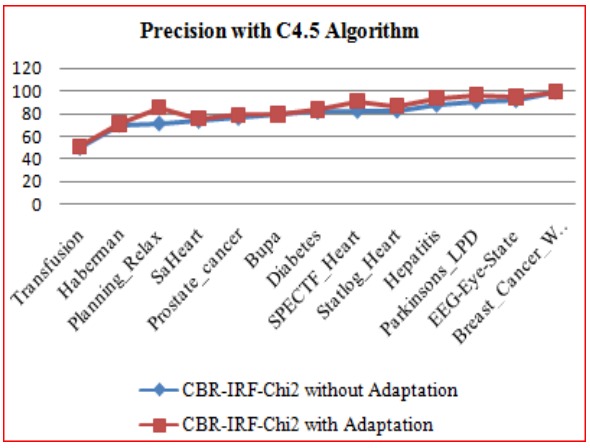
Résultats de Performance :

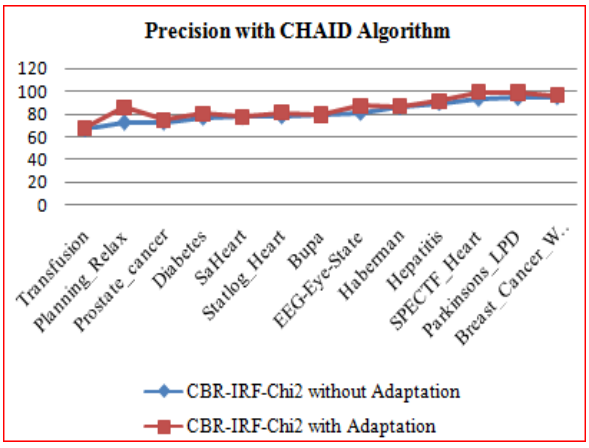
- Le tableau 4.2.9 montre une amélioration de la performance, avec des précisions proches ou supérieures à celles obtenues avec les cas remémorés pour la majorité des bases de données. Cela indique que l'algorithme d'adaptation est efficace.

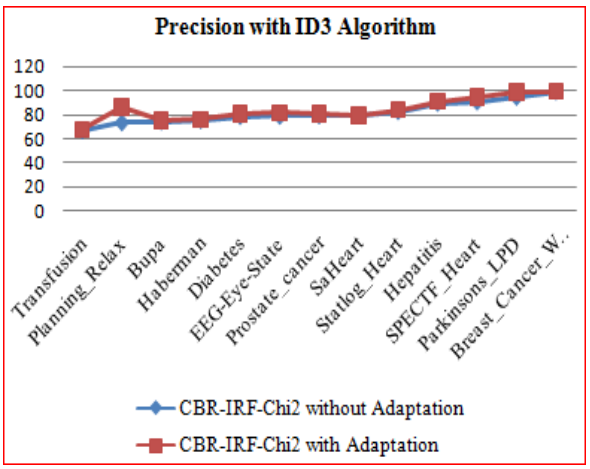
- **Bases avec plusieurs caractéristiques :** Pour des bases telles qu'EEG-Eye-State, Hepatitis, Parkinsons_LPD, Planning_Relax, et SPECTF_Heart, où le nombre de caractéristiques est élevé, la phase d'adaptation améliore les résultats. Les gains en précision varient : 7% pour EEG-Eye-State et Hepatitis, 14% pour Planning_Relax, et 11% pour SPECTF_Heart.

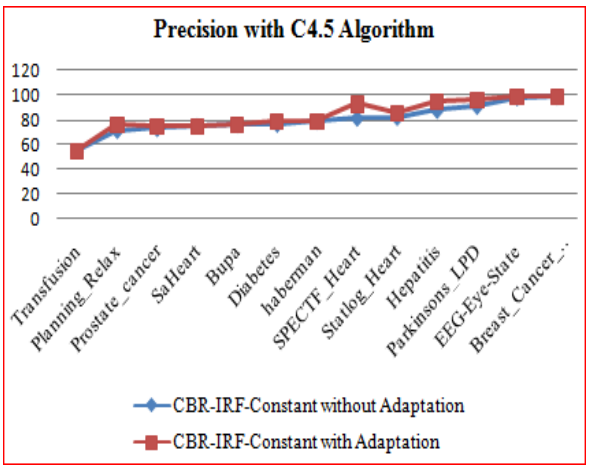
- **Bases avec peu de caractéristiques :** Pour les bases de données avec un nombre limité de caractéristiques, l'approche avec l'algorithme d'adaptation montre également des résultats prometteurs. Les caractéristiques étant pertinentes dès le départ, l'adaptation n'apporte pas de bénéfices significatifs, mais les résultats restent proches de ceux obtenus avec les cas remémorés.

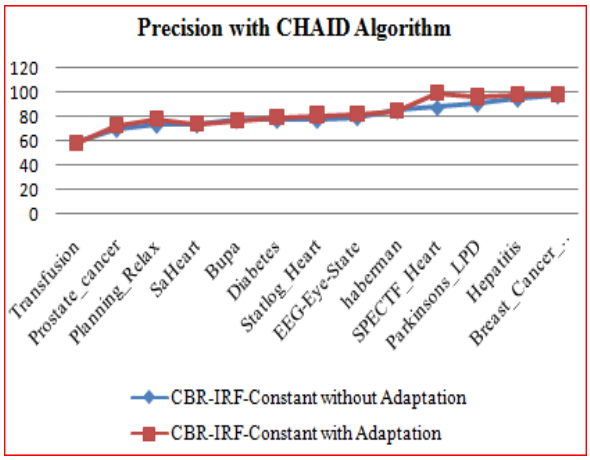
- La figure 4.2.7 démontre que l'algorithme d'adaptation peut offrir des performances optimales, surpassant non seulement les forêts aléatoires classiques, mais aussi les forêts aléatoires modifiées. Cela confirme que la méthode de remémoration guidée par l'adaptation automatique est très efficace en termes de précision.

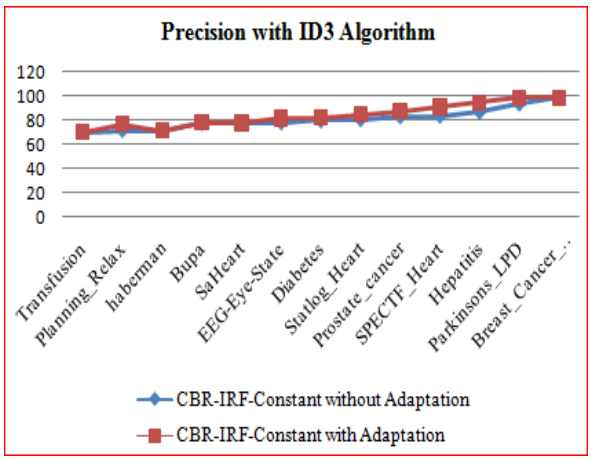

Precision with C4.5 Algorithm
120
100
80
60
40
20
0
Transfusion
Haberman
Planning_Relax
SaHeart
Prostate_cancer
Bupa
Diabetes
SPECTF_Heart
Statlog_Heart
Hepatitis
Parkinsons_LPD
EEG-Eye-State
Breast_Cancer_W..
CBR-IRF-Chi2 without Adaptation
CBR-IRF-Chi2 with Adaptation

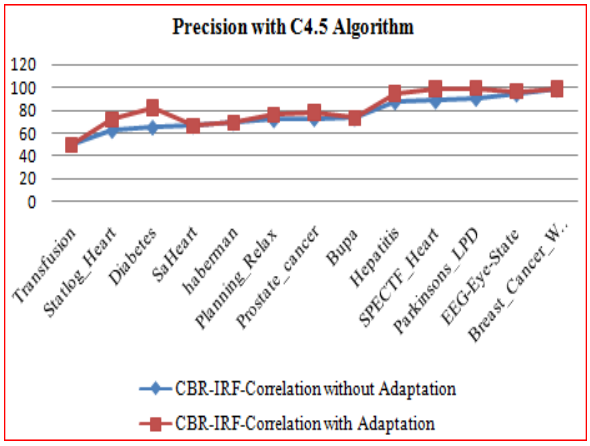

Precision with CHAID Algorithm
120
100
80
60
40
20
0
Transfusion
Planning_Relax
Prostate_cancer
Diabetes
SaHeart
Statlog_Heart
Bupa
EEG-Eye-State
Haberman
Hepatitis
SPECTF_Heart
Parkinsons_LPD
Breast_Cancer_W..
CBR-IRF-Chi2 without Adaptation
CBR-IRF-Chi2 with Adaptation

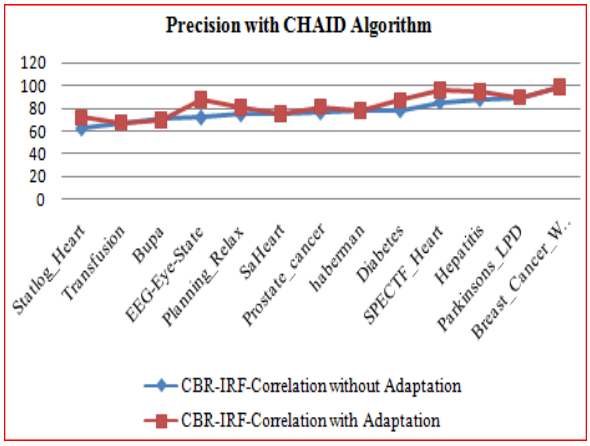

Precision with ID3 Algorithm
120
100
80
60
40
20
0
Transfusion
Planning_Relax
Bupa
Haberman
Diabetes
EEG-Eye-State
Prostate_cancer
SaHeart
Statlog_Heart
Hepatitis
SPECTF_Heart
Parkinsons_LPD
Breast_Cancer_W..
CBR-IRF-Chi2 without Adaptation
CBR-IRF-Chi2 with Adaptation


Precision with C4.5 Algorithm
120
100
80
60
40
20
0
Transfusion
Planning_Relax
Prostate_cancer
SaHeart
Bupa
Diabetes
haberman
SPECTF_Heart
Statlog_Heart
Hepatitis
Parkinsons_LPD
EEG-Eye-State
Breast_Cancer_..
CBR-IRF-Constant without Adaptation
CBR-IRF-Constant with Adaptation


Precision with CHAID Algorithm
120
100
80
60
40
20
0
Transfusion
Prostate_cancer
Planning_Relax
SaHeart
Bupa
Diabetes
Statlog_Heart
EEG-Eye-State
haberman
SPECTF_Heart
Parkinsons_LPD
Hepatitis
Breast_Cancer_..
CBR-IRF-Constant without Adaptation
CBR-IRF-Constant with Adaptation


Precision with ID3 Algorithm
120
100
80
60
40
20
0
Transfusion
Planning_Relax
haberman
Bupa
SaHeart
EEG-Eye-State
Diabetes
Statlog_Heart
Prostate_cancer
SPECTF_Heart
Hepatitis
Parkinsons_LPD
Breast_Cancer_..
CBR-IRF-Constant without Adaptation
CBR-IRF-Constant with Adaptation


Precision with C4.5 Algorithm
120
100
80
60
40
20
0
Transfusion
Statlog_Heart
Diabetes
SaHeart
haberman
Planning_Relax
Prostate_cancer
Bupa
Hepatitis
SPECTF_Heart
Parkinsons_LPD
EEG-Eye-State
Breast_Cancer_W..
CBR-IRF-Correlation without Adaptation
CBR-IRF-Correlation with Adaptation


Precision with CHAID Algorithm
120
100
80
60
40
20
0
Statlog_Heart
Transfusion
Bupa
EEG-Eye-State
Planning_Relax
SaHeart
Prostate_cancer
haberman
Diabetes
SPECTF_Heart
Hepatitis
Parkinsons_LPD
Breast_Cancer_W..
CBR-IRF-Correlation without Adaptation
CBR-IRF-Correlation with Adaptation

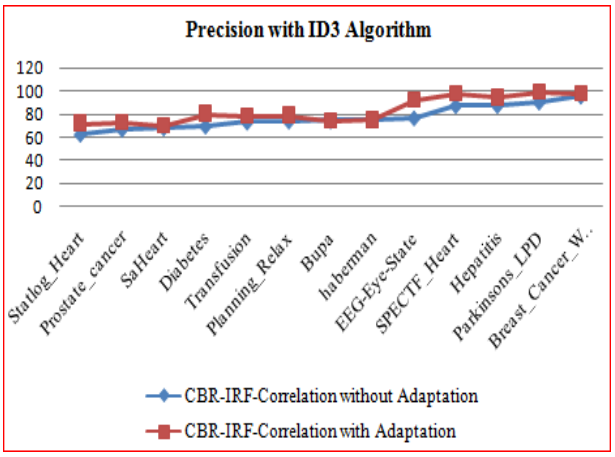


Figure 4.2.7. Performance des méthodes d'apprentissage appliquées avec et sans l'étape de l'adaptation sur 13 bases de données médicales (Tarchoune et al, 2024a)

Tableau 4.2.9. Performance de l'approche proposée avec et sans l'étape d'adaptation sur 13 bases de données médicales (Tarchoune et al, 2024a)

Methodes / Base de données	CBR-IRF-Chi2						CBR-IRF-Constant						CBR-IRF-Correlation					
	Without Adaptation			With Adaptation			Without Adaptation			With Adaptation			Without Adaptation			With Adaptation		
	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3
Breast_Cancer_Wisconsin	98,51	94,44	98,51	98.74	**96.44**	98.71	98,51	97,06	98,48	98.51	**98.21**	98.48	98,48	98,53	95,65	98.48	98.53	**97.25**
Bupa	79,34	79,03	74,05	79.50	79.21	74.85	75,97	76,74	77,52	75.97	76.74	**77.92**	72,99	69,93	74,07	72.99	69.93	74.07
Diabetes	81,34	76,82	78,36	**83.52**	79.92	80.55	76,28	76,92	80,00	78.31	78.87	**82.32**	65,10	77,88	69,43	81.80	**86.98**	80.22
EEG-Eye-State	91,43	81,06	78,60	**94.61**	87.07	81.22	97,78	79,09	78,28	**98.65**	82.08	81.66	94,29	71,70	76,63	**96.27**	87.61	92.20
haberman	70,00	85,71	75,00	71.00	**86.71**	76.00	78,57	84,62	72,00	78.57	**84.62**	72.00	68,97	77,78	75,00	68.97	77.78	75.00
Hepatitis	87,50	89,47	89,19	**93.71**	**91.27**	**90.84**	87,50	94,29	87,18	**94.25**	**98.21**	**94.34**	87,50	87,50	87,50	**94.20**	**94.20**	**94.20**
Parkinsons_LPD	90,20	93,75	94,00	**96.66**	**98.32**	**98.51**	90,74	90,38	94,00	**95.82**	**95.23**	**98.21**	90,57	89,09	90,57	**98.61**	89.08	**98.61**
Planning_Relax	71,19	72,73	73,33	85.22	85.71	**86.11**	71,19	72,88	71,93	76.22	**77.81**	76.92	71,70	74,51	73,77	75.82	**80.21**	78.89
Prostate_cancer	76,19	72,73	78,95	78.75	74.52	**80.27**	72,73	69,57	83,33	74.81	72.88	**87.52**	72,73	76,19	66,67	77.71	**81.19**	72.65
SaHeart	73,48	77,12	79,21	75.62	77.12	79.21	75,00	73,13	77,59	75.00	73.89	77.59	66,24	74,81	67,81	66.24	74.81	**70.20**
SPECTF_Heart	82,14	92,96	90,28	90.77	**98.99**	94.28	81,18	87,50	83,72	**92.72**	**98.99**	**91.27**	88,31	85,19	87,34	**98.21**	**95.71**	**97.31**
Statlog_Heart	82,35	78,05	82,35	**86.71**	81.05	83.57	81,82	76,92	80,95	**85.77**	81.22	84.95	62,22	62,22	62,22	71.57	71.57	**72.22**
Transfusion	50,00	66,67	66,67	51.65	67.67	67.80	54,55	58,82	70,37	54.55	58.28	70.37	50,00	66,67	72.85	50.00	66.67	77.81

Tableau 4.2.10. Taux d'amélioration de l'approche proposée avec et sans l'étape d'adaptation sur 13 bases de données médicales (Tarchoune et al, 2024a)

Methodes / Base de données	CBR-IRF-Chi2						CBR-IRF-Constant						CBR-IRF-Correlation					
	Without Adaptation			With Adaptation			Without Adaptation			With Adaptation			Without Adaptation			With Adaptation		
	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3	C4.5	CHAID	ID3
Breast_Cancer_Wisconsin	98,51	94,44	98,51	+0,23	**+2,00**	+0,20	98,51	97,06	98,48	0,00	**+1,15**	0,00	98,48	98,53	95,65	0,00	0,00	**+1,60**
Bupa	79,34	79,03	74,05	+0,16	+0,18	+0,80	75,97	76,74	77,52	0,00	0,00	**+0,40**	72,99	69,93	74,07	0,00	0,00	0,00
Diabetes	81,34	76,82	78,36	**+2,18**	+3,10	**+2,19**	76,28	76,92	80,00	+2,03	+1,95	+2,32	65,10	77,88	69,43	+16,70	**+9,10**	+10,97
EEG-Eye-State	91,43	81,06	78,60	**+3,18**	+6,01	+2,62	97,78	79,09	78,28	**+0,87**	+2,99	+3,38	94,29	71,70	76,63	**+1,98**	+15,91	+15,57
haberman	70,00	85,71	75,00	+1,00	**+1,00**	+1,00	78,57	84,62	72,00	0,00	0,00	0,00	68,97	77,78	75,00	0,00	0,00	0,00
Hepatitis	87,50	89,47	89,19	**+6,21**	**+1,80**	**+1,65**	87,50	94,29	87,18	**+6,75**	**+3,92**	**+7,16**	87,50	87,50	87,50	**+6,70**	**+6,70**	**+6,70**
Parkinsons_LPD	90,20	93,75	94,00	**+6,46**	**+4,57**	**+4,51**	90,74	90,38	94,00	**+5,08**	**+4,85**	**+4,21**	90,57	89,09	90,57	**+8,04**	+0,71	**+8,04**
Planning_Relax	71,19	72,73	73,33	+14,03	+12,98	**+12,78**	71,19	72,88	71,93	+5,03	**+4,93**	+4,99	71,70	74,51	73,77	+4,12	**+5,70**	+5,12
Prostate_cancer	76,19	72,73	78,95	+2,56	+1,79	**+1,32**	72,73	69,57	83,33	+2,08	+3,31	**+4,19**	72,73	76,19	66,67	+4,98	**+5,00**	+5,98
SaHeart	73,48	77,12	79,21	+2,14	0,00	0,00	75,00	73,13	77,59	0,00	0,76	0,00	66,24	74,81	67,81	0,00	0,00	**+2,39**
SPECTF_Heart	82,14	92,96	90,28	+8,63	**+6,03**	+4,00	81,18	87,50	83,72	**+11,54**	**+11,49**	**+7,55**	88,31	85,19	87,34	**+9,90**	**+10,52**	**+9,97**
Statlog_Heart	82,35	78,05	82,35	**+4,36**	+3,00	+1,22	81,82	76,92	80,95	**+3,95**	+4,30	+4,00	62,22	62,22	62,22	+9,35	+9,35	**+10,00**
Transfusion	50,00	66,67	66,67	+1,65	+1,00	+1,13	54,55	58,82	70,37	0,00	+0,54	0,00	50,00	66,67	72.85	0,00	0,00	+4,96

Tableau 4.2.11. Résultats comparatifs de l'approche proposée par rapport à certains travaux similaires (Tarchoune et al, 2024a)

Methodes	**Base de données**	**Précision**
Hybride CBR, RF (Darabietal, 2014)	Asthma	80%
Hybride CBR, RF (Ayeldeenetal, 2015)	Breast cancer	99%
Hybride CBR, RF (Asimetal, 2019)	Standard blogger dataset	95%
Hybride CBR, RF (SharmaetMehrotra, 2021)	Liver disease	73%
Hybride CBR, RF (AbouDabousetal, 2022)	Pavementmanagement	94%
Notre approche	Breast Cancer_Wisconsin	**98%**
	Bupa	79%
	Diabetes	86%
	EEG-Eye-State	**98%**
	Haberman	86%
	Hepatitis	**98%**
	Parkinsons_LPD	**98%**
	Planning_Relax	86%
	Prostate_cancer	87%
	SaHeart	79%
	SPECTF_Heart	**98%**
	Statlog_Heart	86%
	Transfusion	77%

Les expériences réalisées sur 13 bases de données médicales montrent que le modèle hybride FS-CBR-IRF atteint un haut degré de précision et offre une bonne stabilité en termes de temps de calcul et de précision de classification. Le modèle a prouvé sa fiabilité en s'adaptant à différentes tailles de bases de données et nombres de caractéristiques.

L'approche FS-CBR-IRF surpasse certaines méthodes couramment utilisées et atteint des performances comparables à celles des algorithmes entièrement traitants des systèmes hybrides de raisonnement à partir de cas dans le domaine de la santé.

La combinaison des avantages des sélections de caractéristiques et des classifieurs RàPC et RF a permis d'établir un diagnostic précis pour certaines maladies. Le modèle proposé aide les cliniciens à construire un modèle de prédiction respectueux de la vie privée avec des performances idéales.

L'approche hybride FS-CBR-IRF offre une amélioration significative en termes de taux de classification, temps de calcul, et capacité de discrimination par rapport à l'hybridation de CBR avec RF classique. Elle démontre une grande valeur pour les futures applications de l'intelligence artificielle médicale, permettant aux cliniciens de bénéficier d'un modèle de prédiction précis tout en préservantla vie privée des patients.

4.2.5 Conclusion

Dans cette contribution, nous présentons une nouvelle approche hybride qui fusionne plusieurs méthodes d'apprentissage automatique pour la classification des données médicales. Cette approche combine trois techniques de sélection de caractéristiques, un algorithme amélioré de forêt aléatoire pour optimiser la phase de remémoration et d'adaptation dans un système de raisonnement à partir de cas (RàPC).

Utilisation de techniques de sélection de caractéristiques pour réduire le nombre de caractéristiques et sélectionner les plus pertinentes, améliorant ainsi la base de cas. Cette sélection est réalisée à l'aide de validation croisée pour garantir la pertinence des caractéristiques.

Intégration de l'algorithme de forêt aléatoire amélioré dans la phase de remémoration. Les caractéristiques hautement classées sont utilisées pour entraîner et structurer le modèle, optimisant ainsi la performance de classification.

Application d'une étape d'adaptation guidée par la remémoration pour adapter les cas remémorés si nécessaire. Cette phase vise à améliorer la précision du modèle tout en maintenant une vitesse d'apprentissage efficace.

4.3 Contribution 2: Comparaison d'une approche hybride CBR-RF et CBR-DT pour la classification des données médicales

La classification est un domaine clé dans les algorithmes d'apprentissage automatique, particulièrement pertinent pour les grandes quantités de données médicales disponibles dans les bases de données. Cette contribution se concentre sur le développement d'une approche hybride pour la classification des données médicales, basée sur le raisonnement à partir de cas (RàPC). L'objectif est de modéliser la phase de remémoration du cycle RàPC en utilisant des algorithmes d'arbres de décision (C4.5, RepTree, LMT) et des forêts aléatoires (RF).

L'approche est entraînée et testée sur quarte bases de données médicales à savoir: Wisconsin Breast Cancer, Tyroïde, Hépatites et Breast pathologies. L'importance de cette contribution réside dans la conception et l'implémentation d'un classifieur automatique pour modéliser la phase de remémoration d'un système RàPC.

4.3.1 Composants de l'approche proposée

Différents algorithmes ont été utilisés pour modéliser la phase de remémoration duRàPC. Cette contribution fournit une description détaillée des algorithmes de décision (C4.5, REPTree et LMT), RàPC et les Forêts aléatoires et les évaluer en se basant sur les mesures de performance telles que le taux de classification, l'erreur quadratique.

4.3.1.1 Phase de remémoration du RàPC

La recherche de cas est la première étape du cycle RàPC qui nécessite l'utilisation de mesure de similarité pour trouver des cas similaires au nouvelle cas lorsque la base de cas est trop importante, le calcul de similarité sera couteux en temps de calcul, pour remédier à ce problème, nous avons intégré les arbres de décision et les Forêts aléatoire dans la phase de recherche pour rechercher les cas similaires, proposé et détaillé dans la section 3.1 dans le chapitre1.

4.3.1.2 Arbre de décision

Un arbre de décision est un outil d'aide à la décision reposant sur un ensemble de règles divisant une population de cas en groupe homogènes (Gordon et al, 1984), chaque règle associe une conjonction de tests sur certains attributs avec un groupe, ces règles sont organisés comme un arbre. Divers algorithmes d'arbre de décision (C4.5, LMT, RepTree) ont été utilisés pour la classification dans notre étude.

A. Algorithme C4.5

L'algorithme C4.5 est l'extension de l'algorithme ID3 de l'arbre de décision avec des fonctionnalités supplémentaires telles que la comptabilisation des valeurs manquantes (Karegowda et al, 2012; Mazid et al, 2010)**,** l'élagage des erreurs réduit, la valeur d'attribut continue et la dérivation des règles, etc. L'arbre de décision est une technique supervisée qui construit la classification en arborescence avec le nœud racine, le nœud de branchement et le nœud feuille. L'arbre de décision décompose l'ensemble de données en plusieurs sous-ensembles et construit l'arbre de décision de manière incrémentielle.

B. Algorithme LMT

L'arbre à modèle logistique (LMT) est l'un des classifieurs à arbre de classification. Il utilise une combinaison de méthodes d'apprentissage automatique d'arbre de décision et de régression logistique (KASS, 1980; Landwehr et al, 2005). Dans LMT, les algorithmes de classification et de régression sont utilisés pour élaguer l'arbre pour la classification, tandis que l'algorithme LogitBoost est utilisé pour construire le modèle de régression logistique à chaq ue nœud de l'arbre. Est effectué par le gain d'information de la variante logistique. Pour trouver le nombre d'itérations de LogitBoost, le LMT utilise la validation croisée pour éviter le sur-ajustement. La régression logistique additive de l'ajustement des moindres carrés est utilisée dans l'algorithme LogitBoost pour chaque classe N i comme suit (KASS, 1980) :

$$LN(x) = \sum_{i=1}^{n} \alpha i x i \; + \; \alpha o \qquad (4.3.1)$$

Où n est le nombre de facteurs et αo et αi sont respectivement, le coefficient initial et le coefficient de la i-ème composante du vecteur x.
Dans LMT, les probabilités postérieures des nœuds feuilles sont réalisées en utilisant la méthode de régression logistique linéaire.

$$\mathrm{P}(\mathrm{N}|\mathrm{x}) = \frac{\exp(\mathrm{L}\,\mathrm{N}(\mathrm{x}))}{\sum_{N\prime=1}^{C} exp(LN'(x))} \qquad (4.3.2)$$

Où C est le nombre de classes.

C. Algorithme RepTree

Le RepTree est un arbre de décision formé sur la base du gain d'information. Le site gain d'information est lié à la réduction de la variance. Le RepTree effectue tout fonctionnement de base du C4.5 et effectue également l'élagage en triant les attributs numériques (Mohamed et al, 2012). Cet algorithme fonctionne sur le principe du gain d'information avec l'entropiece qui permet de réduire la variance. Il réduit la complexité des méthodes d'arbre de décision.

L'arbre de décision construit l'arbre par apprentissage supervisé et l'arbre est formé sur la base de l'approche " diviser pour régner ". Il répète la fonction de test de manière récursive jusqu'à ce qu'il soit atteint les nœuds feuilles. L'arbre de décision est utilisé à des fins de classification en raison de sa bonne précision et de sa facilité de traitement (Mohamed et al, 2012)**.** Supposons que A et B soient deux variables distinctes avec les valeurs {a1, ..., an} et {b1, ..., bn}. L'entropie et l'entropie conditionnelle de B sont indiquées dans les équations (4.3.3) et (4.3.4). Ensuite, le gain d'information de A est calculé comme indiqué dans l'équation (4.3.5) :

$$\mathrm{H}(\mathrm{B}) = -\sum_{i=1}^{k} \mathrm{P}(\mathrm{B} = \mathrm{bi}) \log \mathrm{P}(\mathrm{B} = \mathrm{bi}) \qquad (4.3.3)$$

$$\mathrm{H}(\mathrm{B} \setminus \mathrm{A}) = -\sum_{i=1}^{l} \mathrm{P}(\mathrm{A} = \mathrm{ai})\, \mathrm{H}(\mathrm{B} \setminus \mathrm{A} = \mathrm{ai}) \qquad (4.3.4)$$

$$IG(B, A) = H(B) - H(B \setminus A) \qquad (4.3.5)$$

L'élagage peut être effectué de deux manières, à savoir le pré-élagage et le post-élagage. Dans le cas du pré-élagage, l'expansion de l'arbre est arrêtée lorsque le gain d'information dû à la division n'est pas reconnu, tandis que dans le post-élagage, le processus continue jusqu'à ce que tous les nœuds feuilles purs soient trouvés. Les résultats du post-élagage ont meilleurs que ceux du pré-élagage (Kass, 1980).

4.3.1.3 Forêt aléatoire

L'algorithme de la forêt aléatoire est reconnu comme l'une des techniques de classification les plus efficaces, capable de traiter de grandes quantités de données avec une précision élevée. Cette méthode d'apprentissage en ensemble génère un certain nombre d'arbres de décision pendant l'entraînement et fournit la classe la plus fréquente parmi celles prédites par les arbres individuels.

La forêt aléatoire, développée par Breiman, est une approche d'apprentissage qui regroupe plusieurs arbres de décision aléatoires et agrège leurs prédictions en calculant la moyenne. Cette méthode est expliquée en détail dans le chapitre 2.

4.3.1.4 Technique de validation

La validation croisée est une technique utilisée pour évaluer la performance d'un modèle en divisant l'ensemble des données en plusieurs sous-ensembles. Cette méthode permet de tester le modèle de manière plus robuste en utilisant différents échantillons pour l'entraînement et le test. L'une des méthodes de validation croisée les plus couramment utilisées est la validation croisée en k-fold. La procédure de validation croisée en k-fold se déroule comme suit (Latief et al, 2019) :

- Divisez les données en k sous-ensembles de taille égale.
- Utilisez un sous-ensemble comme données de test et les k-1 autres sous-ensembles comme données d'entraînement.
- Répétez ce processus k fois, chaque sous-ensemble servant à tour de rôle de données de test, tandis que les autres servent à l'entraînement.

La précision moyenne de chaque itération est calculée pour obtenir une estimation globale de la performance du modèle. Dans cette étude, nous appliquons une validation croisée à 10 plis, ce qui signifie que les données sont divisées en dix sous-ensembles.

4.3.2 Approche proposée

Vue de l'importance du RàPC dans le diagnostic médical. Nous avonsproposé une approche hybride basée sur le raisonnement à partir de cas qui utilise quatre techniques d'apprentissage supervisé comme technique de recherche, pour trouver les cas les plus similaires dans la bibliothèque de cas (Figure 4.3.1).

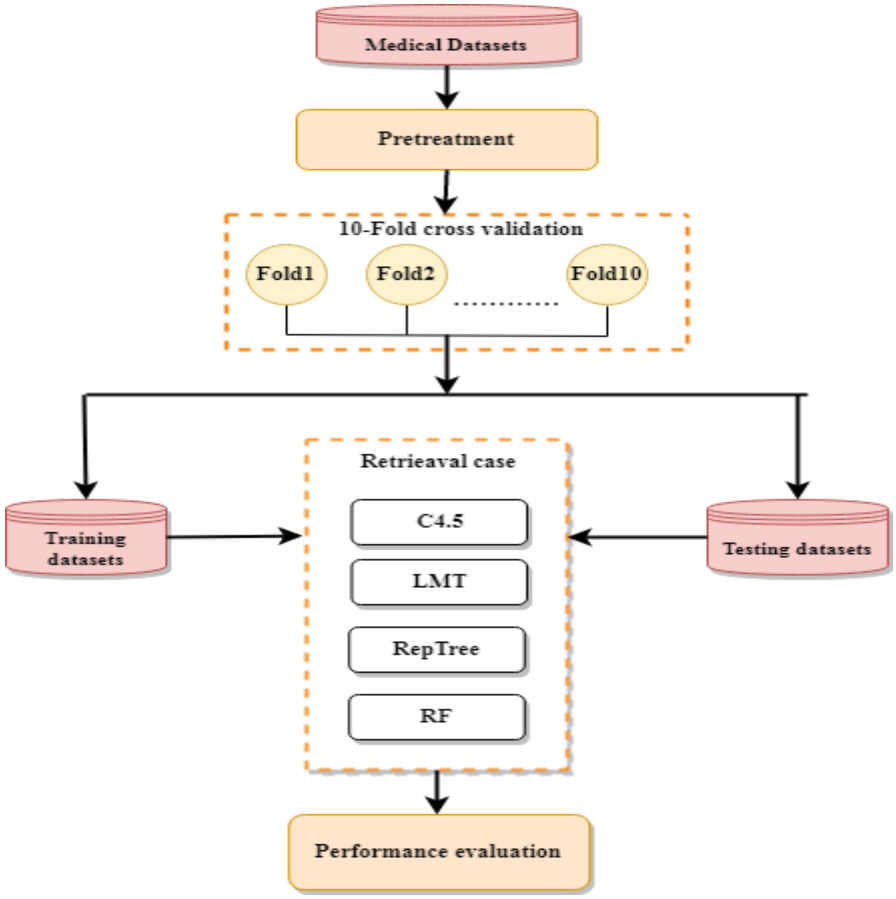


Figure 4.3.1. Architecture générale de l'approche proposée (Tarchoune et al, 2021)

4.3.2.1 Description des bases de données

Dans cette section, nous avons utilisé une base réelle établie de l'hôpital IBN SINA Annaba-Algérie sur les maladies de sein et trois bases de l'UCI.

Tableau 4.3.1. Description de l'ensemble de données (Tarchoune et al, 2021)

Base de données	Type	Taille BDD	Nombre Attribut	Reference
Wisconsin Breast Cancer	Bi-classe	699	10	UCI
Hépatites	Bi-classe	155	20	UCI
Maladie du sein	Multi-classe	100	22	REF
Tyroïde	Multi-classe	215	5	UCI

4.3.2.2 Prétraitement des données

Les performances optimales d'un classifieur automatique dépendent largement de la qualité de la phase de prétraitement. Des données mal prétraitées compromettent la fiabilité du classifieur. Nous avons donc débuté notre section par le prétraitement des données, une étape cruciale. Dans cette partie, nous avons converti notre base de données en un fichier arff (Attribute-Relation File Format), car le logiciel WEKA utilise ce format pour stocker les données (dans notre cas, il s'agit de la base BDD_classe. arff). Notre base initiale était au format Access, il était donc nécessaire de la transformer en format .arff pour l'utiliser dans la classification supervisée. Parmi les transformations possibles, il y a l'élimination des valeurs manquantes, car le classifieur forét aléatoire ne traite pas ces données. Ainsi, pour la classification semi-supervisée, nous avons converti cette base en une base de données numérique pour pouvoir l'utiliser dans l'approche proposée.

Un fichier arff se compose d'une liste d'exemples définis par leurs valeurs d'attributs. Il contient toujours trois types d'informations : un nom pour la base de données, des attributs, et les données. Nous avons donc transformé notre base en un fichier de format .arff.

4.3.2.3 Apprentissage et classification

Dans cette section, nous avons modélisé la remémoration du système RàPC par quatre algorithmes d'apprentissage.

A. Remémoration par l'algorithme C4.5

Nous avons appliqué l'algorithme C4.5 dans la phase de remémoration (algorithme 4.3.1).

Algorithme 4.3.1: CBR-C4.5

Entrée : Base_de_cas (CB)

Sortie : Modélisation_CB_par_C4.5 Trouver une solution pour le cas cible

1. Vérifiez la base de cas.
2. Obtenez le gain d'information pour chaque attribut.
3. Choisissez bmaxb_ {max} bmax comme l'attribut avec le gain d'information maximal.
4. Créez un nœud qui divise selon bmaxb_ {max} bmax.
5. Répétez sur les sous-listes obtenues par la division selon bmaxb_ {max} bmax

   et ajoutez ces nœuds en tant qu'enfants.
6. Assignez la solution résultante au cas cible.

B. Remémoration par l'algorithme LMT

Nous avons appliqué l'algorithme LMT dans la phase de remémoration (algorithme 4.3.2).

Algorithme 4.3.2: CBR-LMT

Entrée : Base_de_cas (CB)
Sortie : Modélisation_CB_par_LMT Trouver une solution pour le cas cible

1. Construisez l'arbre en utilisant les exemples de données d'entraînement.
2. Faites croître l'arbre du modèle logistique en divisant récursivement l'espace des instances.
3. initLogitBoost initialise les probabilités / poids pour l'algorithme LogitBoost.
4. Créez plusieurs arbres de modèles logistiques.
5. Réduction de la taille des arbres (de l'arbre non élagué à l'arbre élagué).
6. Assignez la solution résultante au cas cible.

C. Remémoration par l'algorithme RepTree

Nous avons appliqué l'algorithme RepTree dans la phase de remémoration (algorithme 4.3.3).

Algorithme 4.3.3: CBR-RepTree

Entrée : Base_de_cas (CB)
Sortie : Modélisation_CB_par_RepTree Trouver une solution pour le cas cible

1. Vérifiez la base de cas.
2. Créez plusieurs arbres en utilisant le gain d'information et l'entropie.
3. Si (attributs numériques) : triez tous les attributs numériques.
4. Créez un arbre de décision avec une liste triée.
5. Sinon : créez un arbre de décision avec élagage par erreur.
6. Identifiez le meilleur arbre dans la liste construite.
7. Assignez la solution résultante au cas cible.

D. Remémoration par l'algorithme RF

Nous avons appliqué l'algorithme RF dans la phase de remémoration (algorithme 4.3.4).

Algorithme 4.3.4: CBR-RF

Entrée : Base_de_cas (CB), T le jeu d'apprentissage, L le nombre d'arbres dans la forêt, K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud
Sortie : Forêt_avec_arbres_construits
Modélisation_CB_par_RF Trouver une solution pour le cas cible

1. Pour i de 1 à L faire
2. Tl ← ensemble de bootstrap, dont les données sont générées aléatoirement (avec remplacement) à partir de T
3. arbre ← un arbre vide, c'est-à-dire composé uniquement de sa racine
4. arbre.racine ← RndTree(arbre.racine, Tl, K)
5. forêt ← forêt ∪ arbre
6. Assignez la solution résultante au cas cible

4.3.3 Évaluation de classification

4.3.3.1 Taux de classification (Accuracy)

Cette section présente le taux de classification pour les 04 bases de données par l'application des algorithmes d'apprentissage (CBR-RF, CBR-C4.5, CBR-RepTree, CBR- LMT) (tableau 4.3.2).

Tableau 4.3.2. Taux de classification des méthodes utilisées (Tarchoune et al, 2021)

Algorithmes / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	95,75%	93,99%	92,82%	95,90%
3	96,78%	94,87%	93,85%	96,04%
4	96,05%	94,73%	93,55%	96,48%
5	97,51%	95,46%	94,58%	96,77%
6	96,78%	94,73%	95,46%	96,48%
7	97,21%	95,90%	94,58%	96,77%
8	97,07%	94,14%	94,43%	95,90%
9	97,36%	94,87%	94,72%	96,63%
10	97,22%	96,05%	95,31%	96,77%
11	97,22%	95,75%	95,46%	96,92%
Moyenne	**96,90%**	**95,05%**	**94,48%**	**96,47%**

Wisconsin Breast Cancer

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	93,95%	90,23%	86,97%	94,88%
3	94,88%	88,83%	87,44%	94,41%
4	96,27%	93,48%	92,55%	95,81%
5	93,95%	91,62%	93,02%	97,67%
6	94,41%	91,16%	91,16%	95,81%
7	96,27%	94,88%	91,62%	97,20%
8	95,81%	94,41%	91,16%	96,27%
9	95,81%	93,95%	90,23%	95,81%
10	94,88%	92,09%	92,09%	97,67%
11	94,88%	93,95%	91,62%	97,20%
Moyenne	**95,11%**	**92,46%**	**90,79%**	**96,27%**

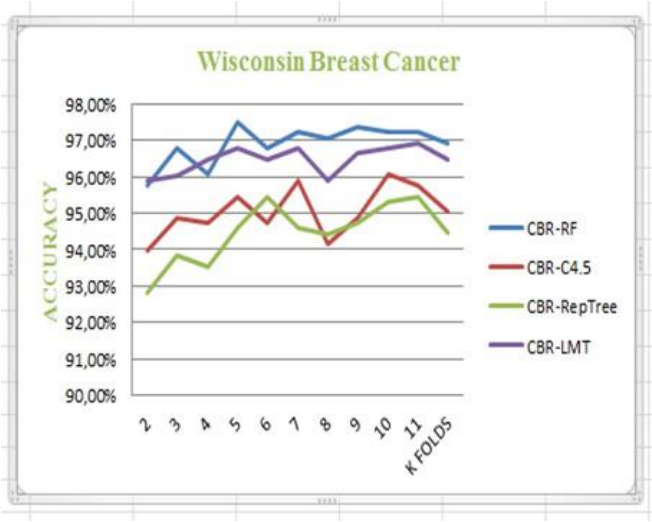
Tyroïde

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	82,58%	81,29%	79,35%	83,87%
3	83,87%	80,64%	79,35%	85,16%
4	83,87%	80,64%	81,29%	83,22%
5	84,51%	80,64%	81,93%	84,51%
6	85,80%	82,58%	79,35%	85,16%
7	79,35%	80,64%	81,93%	81,93%
8	83,87%	76,12%	80,64%	81,93%
9	82,58%	76,77%	77,41%	83,22%
10	85,16%	83,87%	78,70%	83,22%
11	84,51%	81,29%	79,35%	83,22%
Moyenne	**83,61%**	**80,45%**	**79,93%**	**83,54%**

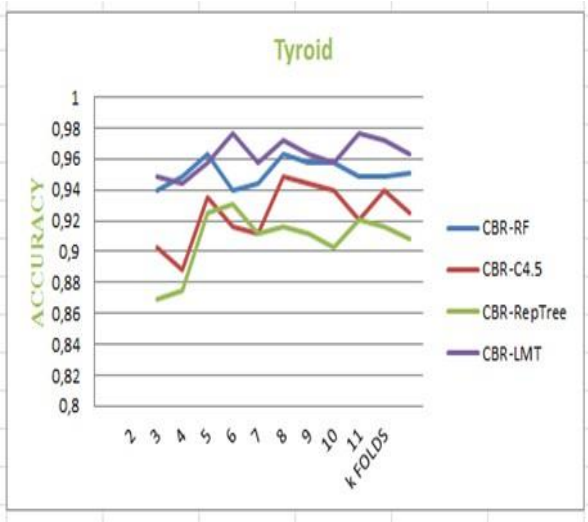
Hepatitis

Algorithme / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	93%	92%	82%	90%
3	95%	95%	89%	94%
4	93%	95%	89%	89%
5	96%	95%	91%	96%
6	94%	95%	88%	93%
7	94%	95%	90%	93%
8	94%	95%	90%	94%
9	94%	95%	88%	95%
10	94%	95%	87%	92%
11	93%	95%	87%	93%
Moyenne	**94%**	**94,70%**	**88,10%**	**92,90%**

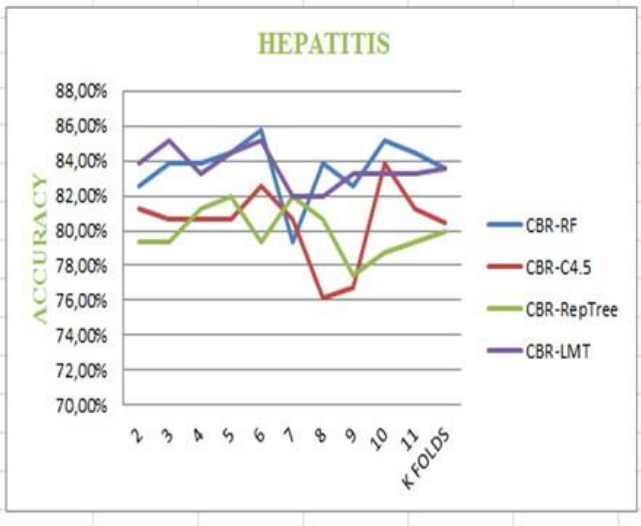
Breast pathologies



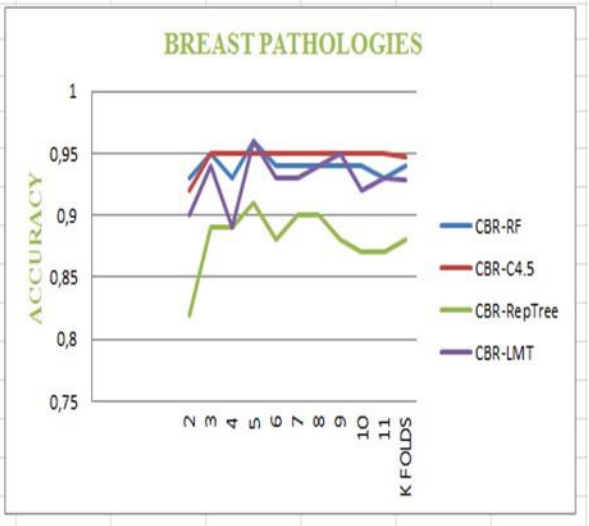
Wisconsin Breast Cancer



Tyroïde



Hepatitis



Breast pathologies

Figure 4.3.2. Taux de classification des méthodes utilisées (Tarchoune et al, 2021)

4.3.3.2 Erreur quadratique moyenne

Par l'application des algorithmes d'apprentissage (CBR-RF, CBR-C4.5, CBR-RepTree, CBR-LMT) sur les quatre bases de données, nous avons obtenus l'erreur quadratique moyenne présentée dans les tableaux (tableau 4.3.3).

Tableau 4.3.3. Erreur quadratique des méthodes utilisées (Tarchoune et al, 2021)

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	0,1571	0,1709	0,2456	0,2063
3	0,1473	0,1428	0,1932	0,1519
4	0,1491	0,1426	0,1878	0,1979
5	0,1374	0,1427	0,18	0,1344
6	0,142	0,1425	0,2006	0,1648
7	0,141	0,1436	0,1878	0,1572
8	0,1368	0,1419	0,1938	0,1429
9	0,1387	0,1422	0,1982	0,1405
10	0,1372	0,142	0,2069	0,1553
11	0,1427	0,142	0,2114	0,1573
Moyenne	**0,14293**	**0,14532**	**0,20053**	**0,16085**

Wisconsin Breast Cancer

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	0,1571	0,1709	0,2456	0,2063
3	0,1473	0,1428	0,1932	0,1519
4	0,1491	0,1426	0,1878	0,1979
5	0,1374	0,1427	0,18	0,1344
6	0,142	0,1425	0,2006	0,1648
7	0,141	0,1436	0,1878	0,1572
8	0,1368	0,1419	0,1938	0,1429
9	0,1387	0,1422	0,1982	0,1405
10	0,1372	0,142	0,2069	0,1553
11	0,1427	0,142	0,2114	0,1573
Moyenne	**0,14293**	**0,14532**	**0,20053**	**0,16085**

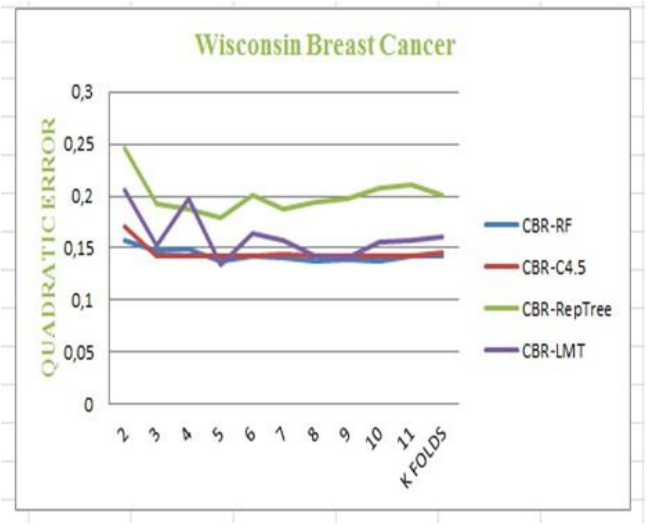
Thyroide

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	0,1685	0,2435	0,2744	0,1614
3	0,1708	0,2626	0,2792	0,1692
4	0,1468	0,1963	0,2178	0,1455
5	0,1475	0,2281	0,2126	0,1178
6	0,1599	0,229	0,2325	0,141
7	0,1403	0,1818	0,2255	0,1037
8	0,1435	0,1817	0,2319	0,1249
9	0,1398	0,193	0,2477	0,1221
10	0,148	0,2191	0,2258	0,1136
11	0,1523	0,196	0,2277	0,1151
Moyenne	**0,15174**	**0,21311**	**0,23751**	**0,13143**

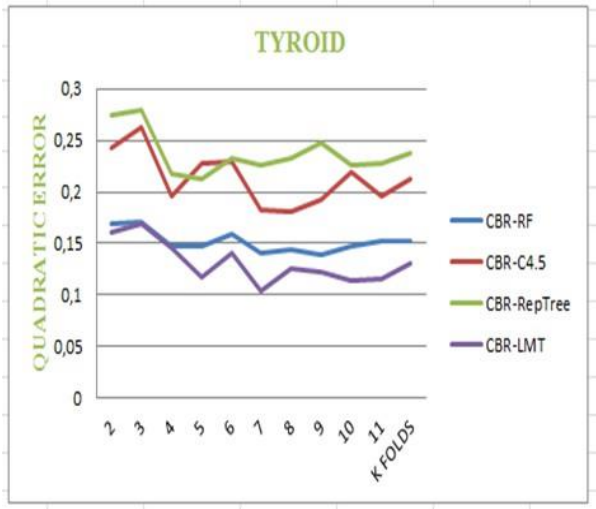
Hépatites

Algorithmes / Test / k-folds	CBR-RF	CBR-C4.5	CBR-RepTree	CBR-LMT
2	0,3422	0,3981	0,4048	0,3657
3	0,3307	0,4151	0,3939	0,3253
4	0,3304	0,4025	0,3828	0,3528
5	0,3417	0,4064	0,3868	0,3447
6	0,3317	0,382	0,3982	0,312
7	0,346	0,4066	0,382	0,3541
8	0,3334	0,4443	0,3931	0,3645
9	0,3424	0,4342	0,4117	0,3578
10	0,3321	0,363	0,4067	0,3437
11	0,3296	0,394	0,406	0,3546
Moyenne	**0,33602**	**0,40462**	**0,39660**	**0,34752**

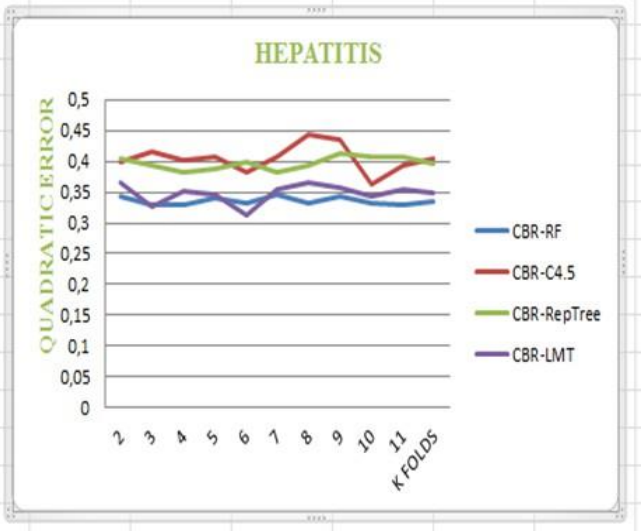
Breast pathologies



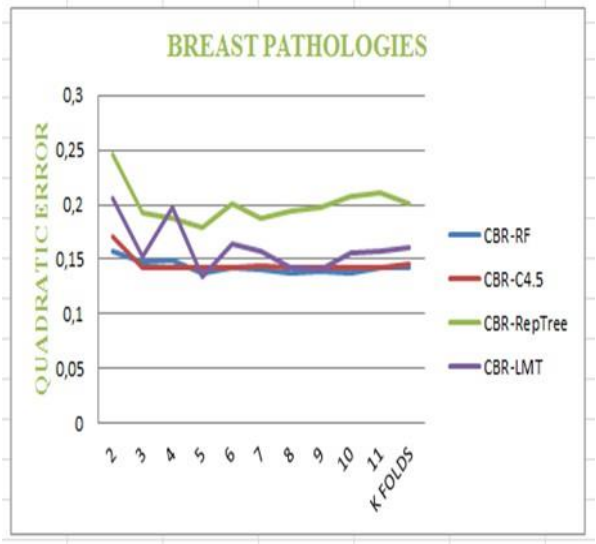
Wisconsin Breast Cancer



Tyroïde



Hepatitis



Breast pathologies

Figure 4.3.3. Erreur quadratique des méthodes utilisées (Tarchoune et al, 2021)

4.3.4 Résultats et discussions

Nous présentons les résultats obtenus de l'approche proposée en termes de taux de classification et d'erreur quadratique moyenne. D'après le tableau 4.3.2 et la figure 4.3.2, nous constatons que :

Les meilleurs taux de classification sur 10 tests effectués sur la base de données du cancer du sein et Wisconsin sont obtenus avec l'algorithme CBR-RF, avec un taux de classification supérieur à 92%. En comparant les taux de classification maximaux des quatre algorithmes sur la même base avec k=5, nous remarquons que l'algorithme CBR-RF atteint un taux de classification de 97.51%, suivi par l'algorithme CBR-LMT à 96.77%. Les autres algorithmes présentent des performances inférieures. Le tableau 4.3.3 montre que l'algorithme CBR-RF donne la plus faible erreur quadratique moyenne avec k=8, confirmant ainsi la performance de cet algorithme pour modéliser la phase de remémoration

Les meilleurs résultats pour la base de données de la thyroïde sont obtenus avec l'algorithme CBR-LMT. En comparant les taux de classification illustrés dans la figure 4.3.4 avec k=5, nous observons que le taux de classification de l'algorithme CBR-LMT est de 97.67%, suivi par l'algorithme CBR-RF avec 93.95%. Le tableau 4.3.3 montre que l'erreur quadratique moyenne reste presque stable pour les algorithmes CBR-LMT et CBR-RF à partir de k=9, indiquant que l'utilisation d'un algorithme de classification dans la phase de remémoration réduit considérablement l'erreur quadratique moyenne.

Les meilleurs taux de classification sur 10 tests effectués sur la base de données des hépatites sont obtenus avec l'algorithme CBR-RF. En comparant les taux de classification avec k=6, nous remarquons que le taux de classification de l'algorithme CBR-RF est de 85.80%, suivi par l'algorithme CBR-LMT à 85.16%. Le tableau 4.3.2 montre que l'algorithme CBR-LMT donne la plus faible erreur quadratique moyenne avec 0.1037.

D'après les figures 4.3.4 et 4.3.5, les algorithmes CBR-RF et CBR-LMT donnent les meilleurs résultats de classification et les plus faibles erreurs quadratiques sur la base de données des pathologies mammaires. Après évaluation des algorithmes sur différentes bases de données, nous concluons que les algorithmes CBR-RF et CBR-LMT sont performants. L'amélioration de la précision de la classification par rapport à un arbre de décision unique est due à l'utilisation de plusieurs arbres de décision générés à partir d'un ensemble de données, chaque arbre votant pour la décision finale.

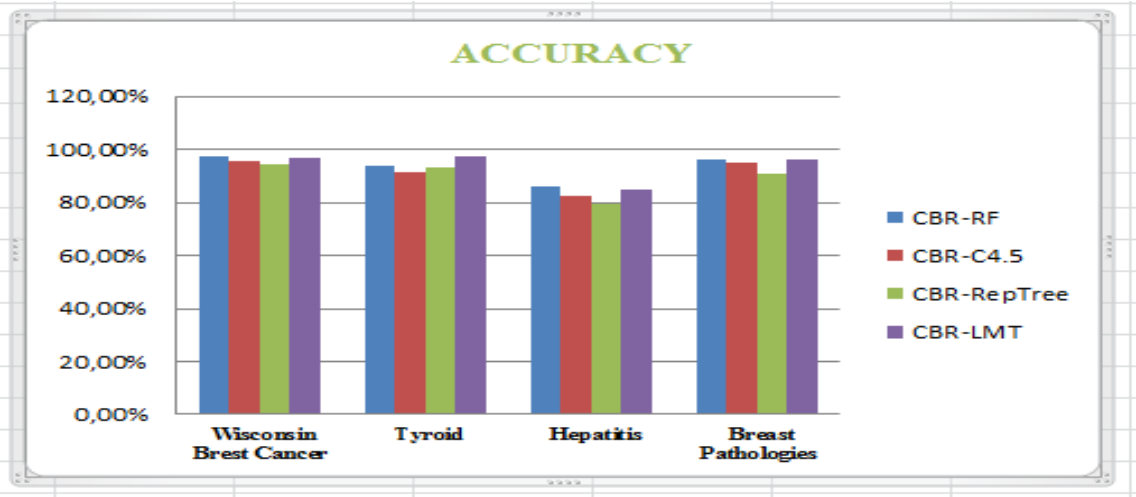


Figure 4.3.4. Comparaison entre les taux de classifications des algorithmes implémentés

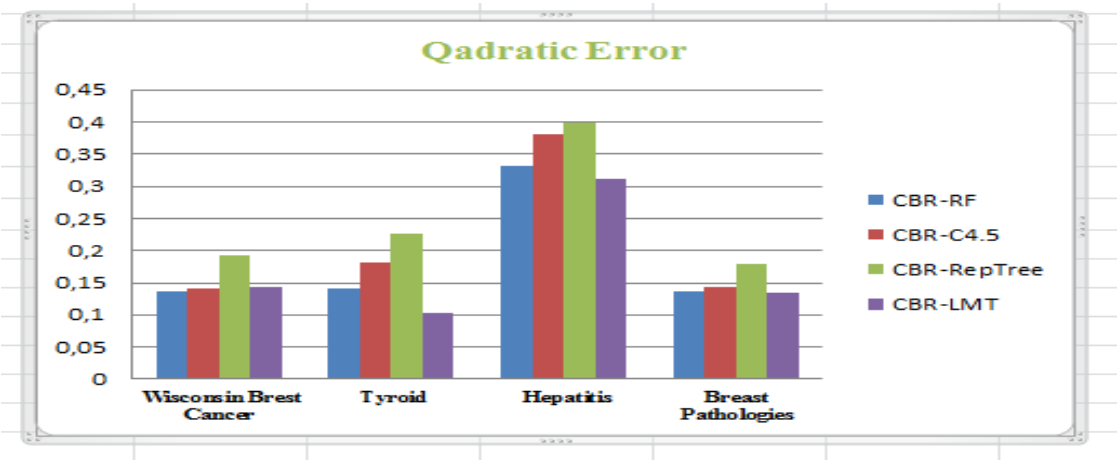


Figure 4.3.5. Comparaison entre les erreurs quadratiques moyennes des algorithmes

Les résultats présentés dans le tableau 4.3.4 indiquent que la combinaison d'un raisonnement à partir de cas et les foréts aléatoires est une direction prometteuse pour une meilleure performance de classification.

Tableau 4.3.4. Les résultats comparatifs de l'approche proposée contre les autres techniques d'apprentissage (Tarchoune et al, 2021)

Méthodes	**Travaux connexes**	**Précision (%)**
CBR-DT	(Quellec et al, 2007)	79.5
	(Rong-Holin, 2009)	90
	(Abdul et al, 2011)	92
CBR-RF	(Darabi et al, 2014)	80
	(Ayeldeen et al, 2015)	95
	(Asim et al, 2019)	99
Notre approche		
CBR-C4.5		95
CBR-LMT		**96**
CBR-Reptree		94
CBR-RF		**97**

4.3.5 Conclusion

Dans cette contribution, nous avons proposé une approche hybride basée sur le raisonnement à partir de cas (RàPC), pour objectif de modéliser la phase de remémoration à l'aide des classifieurs arbre de décision et forêts aléatoires. Nous avons étudié la performance de quatre algorithmes : CBR-RF, CBR-C4.5, CBR-RepTree, et CBR-LMT, pour la classification de quatre bases de données médicales : Wisconsin Breast Cancer, Thyroïde, Hépatites, et maladies du sein.

Les résultats de simulation montrent que les meilleurs taux de classification sont obtenus avec les algorithmes CBR-RF et CBR-LMT. Ces deux algorithmes affichent également les plus faibles taux d'erreur quadratique moyenne. Les expériences réalisées confirment que la stratégie proposée améliore considérablement la performance de classification.

4.4 Proposition d'une forêt aléatoire améliorée basée sur la sélection et la pondération des caractéristiques pour la remémoration des cas dans le système RàPC

Dans cette section, afin d'améliorer la performance de la phase de remémoration du système RàPC, nous avons utilisé les forêts aléatoires de trois manières différentes : forêt aléatoire classique (CRF), forêt aléatoire avec sélection de variables (RF_FS), où nous avons sélectionné les attributs les plus importants et supprimé les attributs les moins pertinents, et forêt aléatoire pondérée (WRF), où nous avons pondéré les attributs les plus importants en leur attribuant plus de poids. Nous avons testé nos trois algorithmes, CBR-CRF, CBR-RF_FS et CBR-WRF, sur 11 bases de données médicales et avons comparé les résultats obtenus.

4.4.1 Architecture générale de l'approche proposée

L'objectif de cette contribution est d'améliorer les forêts aléatoires pour modéliser la phase de remémoration du système RàPC. Nous avons utilisé l'algorithme RF sur plusieurs maladies à l'aide de trois classifieurs supervisés : RF classique, RF avec sélection d'attributs, et RF pondéré. Nous démontrons l'efficacité de ces trois techniques sur onze bases de données médicales.

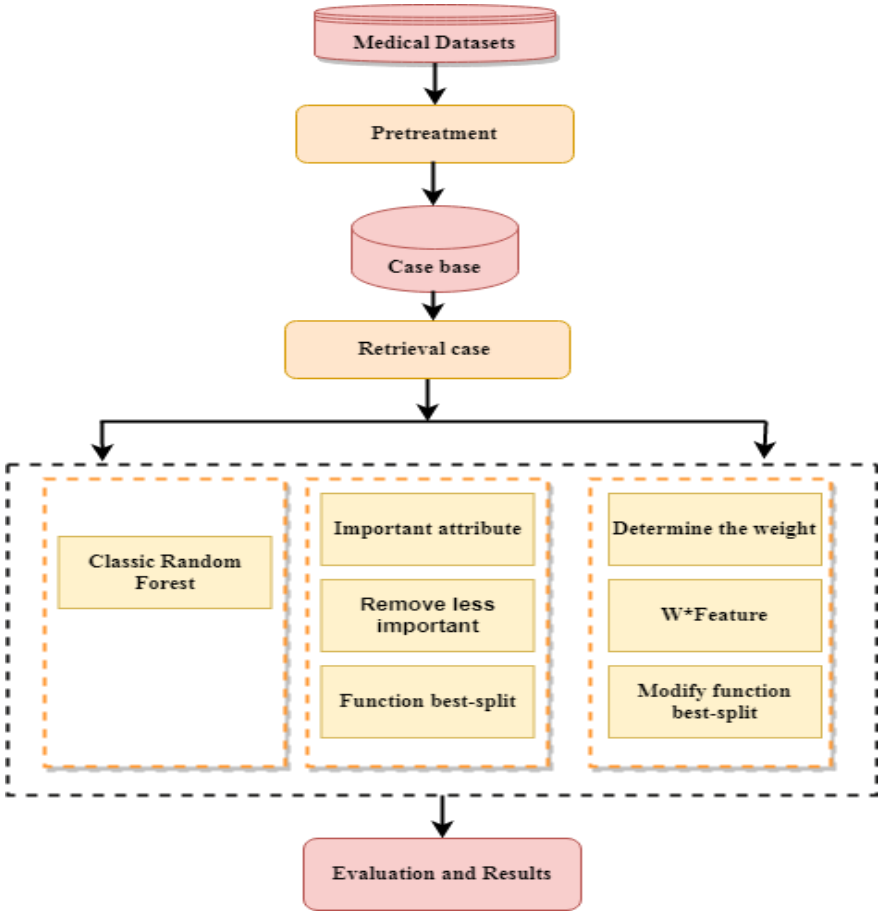


Figure 4.4.1. Architecture générale de l'approche proposée (Tarchoune et al, 2022)

4.4.2 Ensemble de données

Dans le tableau 4.4.1, nous décrivons les onze bases de données utilisées dans notre contribution. Parmi ces onze bases trois sont de Kaggle, un est d'Algérie et sept sont de l'Université de California Irvine (UCI) (tableau 4.4.1).

Tableau 4.4.1. Description des 11 bases utilisées (Tarchoune et al, 2022)

Base de données	Type	Taille BDD	Nbr Attribut	Référence
Pima	Bi-classe	768	8	UCI
Stalogheart	Bi-classe	303	14	UCI
Lung cancer	Bi-classe	59	7	Kaggle
Hepatite	Bi-classe	155	20	UCI
Maladie du sein	Multi-classe	100	28	Algérie
Alzheimer	Multi-classe	354	15	UCI
Eeg-Eye-state	Bi-classe	1498	15	Kaggle
Transfusion du sang	Bi-classe	748	5	Kaggle
Dermatologie	Multi-classe	366	35	UCI
Prostate cancer	Bi-classe	100	10	UCI
Haberman	Bi-classe	306	4	UCI

4.4.3 Modèles de classification

4.4.3.1 Forêt aléatoire classique

Nous avons appliqué l'algorithme de forêt aléatoire classique pour modéliser la phase de remémoration par l'algorithme 4.4.1.

Algorithme 4.4.1: Forêt aléatoire classique (CRF)

Entrée: T l'ensemble d'apprentissage L le nombre d'arbres dans la forêt

K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud

Sortie: forêt l'ensemble des arbres qui composent la forêt construite

1. **Pour** i de1à L **faire**
2. T_l←ensemble bootstrap, dont les données sont tirées aléatoirement (avec remise) de T
3. arbre←un arbre vide, i.ecomposé des aracines uniquement
4. arbre.racine← arbreTree(arbre.racine,T_l,K)
5. forêt ← forêt ∪ arbre

Retour : forêt

L'algorithme suivant présente la construction de l'arbre dans la forêt aléatoire (algorithme 4.4.2).

Algorithme 4.4.2: ArbreTree

Entrée: n le nœud courant
T l'ensemble des données associées au nœud n
K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud
Sortie: n le même nœud, modifié par la procédure

1. **si** n n'est pas une feuille **alors**
2. C←K caractéristiques choisies aléatoirement
3. **Pour tout** A∈C **faire**
4. Procédure ID3 pour la création et l'évaluation (entropy) du partitionnement produit par A, en fonction de T
5. partition← partition qui optimise l'entropy
6. n. ajouter fils (partition)
7. **pour tout** fils ∈ n.noeudFils **faire**
8. RndTree (fils, fils. donnes, K)

Retour n

4.4.3.2 Forêt aléatoire avec sélection des attributs

Dans cette sous-section, nous utilisons le même algorithme que forêt aléatoire classique. Nous supprimons les attributs les moins importants de chaque base de données et nous faisons la classification après une consultation avec l'expert du domaine qui est un médecin de la médecine interne de l'hôpital Beni Messous d'Alger. Il nous a défini les attributs les plus importants (algorithme 4.4.3).

Algorithme 4.4.3: RF_FS proposé

Entrée: T l'ensemble d'apprentissage L le nombre d'arbre dans la forét

K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud

Sortie: forêt l'ensemble des arbres qui composent la forêt construite

1. **Pour** i de1à L **faire**
2. Tl←ensemble bootstrap, dont les données sont tirées aléatoirement (avec remise) DeT
3. arbre←un arbre vide, / composé de sa racine uniquement
4. arbre.racine←RndTree(arbre.racine,Tl,K)
5. forêt ← forêt ∪arbre

Retour forêt

4.4.3.3 Forêt aléatoire pondéré

Dans cette sous-section, nous avons attribué des poids W aux attributs les plus importants selon une fonction

$$W = \frac{\text{nbat}}{\text{nbT}} \qquad (4.4.1)$$

Ou nbat est le nombre des attributs les plus importants et nbT est le nombre des attributs de l'ensemble T

Fonction poids(W) proposé

Entrée: T l'ensemble d'apprentissage atles attributs les plus importants

Sortie : w le poids des attributs

1. nb_at← nombre des attributs les plus importants
2. nb_T← nombre des attributs de l'ensemble T
3. p← nbat/nbT nombre attributs plus important / nombre attributs total
4. **pour i** de 1 a nbT **faire**
5. **Si** i est de *at* **faire**
6. *w[i]* ←p
7. alors *w[i]* ←1

Nous avons modifié notre algorithme CRF (algorithme 4.4.1) au niveau de la fonction determine_best_split où nous avons multiplié l'entropie avec le poids correspondant à chaque attribut.

Algorithme 4.4.4: WRF proposé

Entrée: T l'ensemble d'apprentissage

L le nombre d'arbres dans la forêt

K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud

w Poids des attributs

Sortie: forêt l'ensemble des arbres qui composent la forêt construite

1. **pour** i de1à L **faire**
2. T_l←ensemble bootstrap, dont les données sont tirées aléatoirement (avec remise) de T
3. arbre←un arbre vide, / composé de sa racine uniquement
4. arbre.racine←RndTree(arbre.racine,T_l,K,w)
5. forêt ← forêt ∪ arbre
6. **retour** forêt

Algorithme 4.4.5: ArbreTree

Entrée: n le nœud courant

T l'ensemble des données associées au nœud n

K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud
w poids des attributs

Sortie: n le même nœud, modifié par la procédure

1. **si** n n'est pas une feuille **alors**
2. potential_splits=get_potential_splits(T)
3. split_column,split_value=determine_best_split (T,potential_splits,w)
4. Calculer l'entropy de chaque attribut et Pondéré l'entropy de chaque attribut et sélectionnée l'entropy le plus optimisé pour la partition: data_below, data_above = split_data (data, split_column, split_value)
5. n. ajouter fils (data_below)
6. n. ajouter fils (data_above)
7. **Pour tout** fils∈ n. noeudFils **faire**
8. RndTree (fils, fils. donnes, K)
9. **retour**n

4.4.4 Mesures d'évaluation de la classification

Choix de la variable de segmentation

Entropie d'une seule variable ayant deux catégories, entropie varie entre 0 et 1

$$E(T) = -\sum_{i=1}^{k} p_i \log(p_i) \qquad (4.4.2)$$

Où pi est la probabilité qu'un élément de T.

$$E(T,x) = \sum_{i=1}^{k} p_i E(T) \qquad (4.4.3)$$

On commence le split (division) par la variable ayant la plus grande valeur en gain d'information ou la plus petite valeur d'entropie globale.

Matrice de confusion

Une Confusion Matrix est un résumé des résultats de prédictions sur un problème de classification (Tharwat, 2018). Les prédictions correctes et incorrectes sont mises en lumière et réparties par classe. Les résultats sont ainsi comparés avec les valeurs réelles.

	Positif	**Négatif**
Vrai	Vrai Positif (VP)	Vrai Négatif (VN)
Faux	Faux Positif (FP)	Faux Négatif (FN)

_VP: Vrai Positif: nombre de positifs classés positifs.
_VN: Vrai Négatif: nombre de négatifs classés négatifs.
_FP: Faux Positif: nombre de négatifs classés positifs.
_FN: Faux Négatif: nombre de positifs classés négatifs.

Sensibilité

C'est la capacité de donner un résultat positif quand la maladie est présente.
Sensibilité=VP/nombre total de positifs (=VP/nombre de oui prédits =VP/(VP+FN))

Spécificité

C'est la capacité de donner un résultat négatif quand la maladie est absente.
Spécificité=VN/nombre total de vrais négatifs de la population (=VN/non véritables =VN/tous les vrais négatifs de la population=VN/ (VN+FP))

Taux de classification

C'est le nombre d'exemples bien classés, en valeur absolue, puis en pourcentage du nombretotal d'exemples.
Taux de classification = (VP+VN)/total

Taux d'erreur

Est la proportion des mals classés, il estime la probabilité de mal classer un individu pris au hasard dans la population.
Taux d'erreur= (FP+FN)/total

4.4.5 Expériences et analyses

L'objectif de la classification supervisée est de définir des règles permettant de classer des instances dans des classes en se basant sur des variables qualitatives caractérisant ces instances. Il existe plusieurs méthodes de classification supervisée telles que SVM, arbre de décision, forêt aléatoire, réseaux de neurones, réseaux bayésiens, et k-plus proche voisin (KNN), etc. Nous avons choisi la méthode de forêt aléatoire. La qualité de la classification dépend du taux de classification, c'est-à-dire qu'une bonne classification correspond à un bon taux de classification (une bonne précision). Notre objectif est de maximiser le taux de classification ou de minimiser le taux d'erreur.

Le tableau 4.4.2 contient les attributs les plus importants et leurs poids (W).

Tableau 4.4.2. Les attributs les plus importants et leurs Poids W (Tarchoune et al, 2022)

Base de données	Les attributs les plus importants	Pondéré(WRF) Poids W
Diabete	Glucose, DiabetesPedigreeFunction, BMI, Age	0.5
Alzheimer	CDR","MMSE","Age	0.25
Heart	'age','sex','cp','trestbps','chol','fbs'	0.46
Hepatite	'age','sex','steroids','antivirals'	0.79
Maladiedu sein	Imagerie: mammographie, opacite, clarte, rien_a_signaler, forme, taille, contours, homogeinite, microcalcification, dipistage Biologique:cytologie, biopsie	0.45
Lung Cancer	Smokes, AreaQ, Alkhol	0.75
EEG	AF3,F7,F3,FC5,T7,P7,01,02,P8,T8,FC6,F4,F8,AF4,	1.0
Prostate	Imagerie:radius,texture,perimeter,area,smoothness, compactness, symmetry, fractal_dimension	0.89
Transfusion	Recency, Frequency, Monetary, Time	1.0
Haberman	Number_of_positive_axillary_nodes_detected	0.33
Dermato	Clinique : erythema, scaling, definite_borders, itching, koebner_phenomenon,polygonal_papules,follicular_papules, oral_mucosal_involvement, knee_and_elbow_involvement, scalp_involvement, family_history, melanin_incontinence, eosinophils_in_the_infiltrate, pnl_infiltrate, fibrosis_of_the_papillary_dermis, exocytosis ,acanthosis, hyperkeratosis, parakeratosis, clubbing_of_the_rete_ridges, elongation_of_the_rete_ridges, thinning_of_the_suprapapillary_epidermis, spongiform_pustule, munro_microabcess, focal_hypergranulosis, disappearance_of_the_granular_layer, vacuolisation_and_damage_of_basal_layer, spongiosis, saw-tooth_appearance_of_retes, follicular_horn_plug, perifollicular_parakeratosis, inflammatory_monoluclear_inflitrate,band-like_infiltrate,	0.97

Expérimentation 1

Le tableau 4.4.3 contient les taux de classification d'un RàPC avec un seul arbre de décision (ID3) et un RàPC avec une forêt aléatoire classique.

En comparant les résultats du système RàPC utilisant un arbre de décision avec ceux utilisant une forêt aléatoire, laquelle contient 100 arbres de décision pour chaque base de données, nous remarquons que le RàPC avec la forêt aléatoire fournit de meilleurs résultats par rapport à un seul arbre de décision pour presque toutes les bases de données.

Tableau 4.4.3. Les taux de classification de l'arbre et la forét (Tarchoune et al, 2022)

Bases de données	CBR-Arbre dedécision	CBR-Forêt aléatoire classique
Diabete	0.69	0.77
Hépatite	0.67	0.67
Heart	0.9	0.9
Alzheimer	0.92	0.92
Maladie du sein	1	0.97
Lung Cancer	1	1
EEG	0.85	0.93
Transfusion	0.76	0.77
Prostate cancer	0.71	0.75
Dermato	0.83	0.97
Haberman	0.64	0.73

Expérimentation 2

Dans cette expérimentation, nous comparons la performance (taux de classification, spécificité, sensitivité et taux d'erreur) de trois algorithmes utilisés.

Le premier algorithme, CBR-CRF, intègre une forêt aléatoire classique simple dans le système RàPC. Le deuxième, CBR-RF-FS, est une hybridation du système RàPC avec une forêt aléatoire classique utilisant uniquement les attributs les plus importants d'une base de données (selon un médecin). Le troisième, CBR-WRF, utilise tous les attributs mais pondère les attributs les plus importants.

Tableau 4.4.4. La performance du CBR-RF proposée sur les différents classifieurs (Tarchoune et al, 2022)
(Les meilleurs résultats sont en gras)

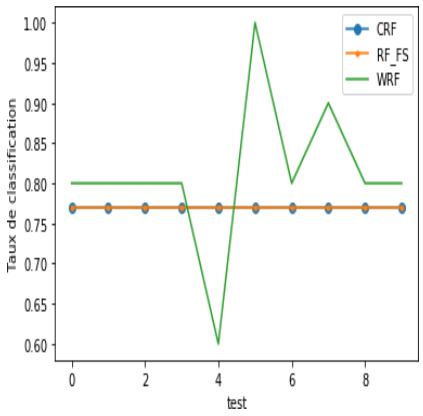
Base de données	CBR-CRF				CBR-RF-FS				CBR-WRF			
	TC	Se	Sp	TE	TC	Se	Sp	TE	TC	Se	Sp	TE
Diabète	0.77	0.71	0.71	0.23	0.77	0.74	0.74	0.23	**0.81**	**0.778**	**0.778**	**0.19**
Hépatite	0.67	nan	nan	0.33	0.33	0.25	0.25	0.67	**0.91**	nan	nan	**0.09**
EEG	0.93	0.93	0.93	0.07	**0.94**	**0.94**	**0.94**	**0.06**	0.76	0.74	0.74	0.24
Heart	0.7	0.61	0.71	0.3	0.8	0.8	0.8	0.2	**0.83**	**0.81**	**0.81**	**0.17**
Alzheimer	0.92	0.9	0.9	0.08	**1**	**1**	**1**	**0**	0.91	0.96	0.96	0.09
Lung cancer	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0.0**
Maladie de sein	0.88	0.94	0.94	0.12	0.83	0.92	0.92	0.17	**0.91**	**0.95**	**0.95**	**0.09**
Transfusion du sang	0.76	0.62	0.62	0.24	**0.8**	**0.7**	**0.7**	**0.2**	0.77	nan	nan	0.23
Prostate cancer	0.75	0.5	0.5	0.25	0.75	0.5	0.5	0.25	**0.85**	**0.85**	**0.85**	**0.11**
Haberman	0.73	0.64	0.64	0.27	**0.82**	**0.9**	**0.9**	**0.18**	0.73	nan	nan	0.27
Dermatology	**0.97**	**0.98**	**0.98**	**0.03**	**0.97**	**0.98**	**0.98**	**0.03**	0.92	0.96	0.96	0.08

Dans le tableau 4.4.4, nous présentons les résultats obtenus de l'approche proposée en termes de précision, spécificité, sensitivité et taux d'erreur.

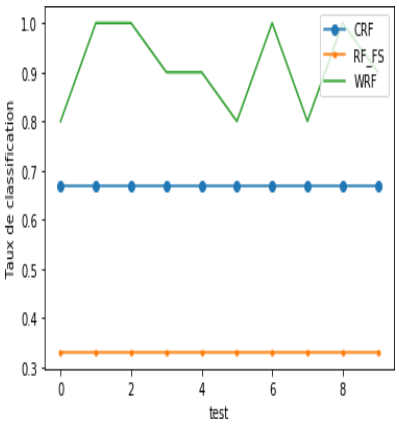
Dans les systèmes médicaux, la spécificité est une mesure importante pour juger l'efficacité du processus de diagnostic (d'après un médecin). Dans notre contribution, nous avons atteint des valeurs significatives de spécificité dans la plupart des bases de données : EEG (0.94), Alzheimer (1), Maladie du sein (0.95), Cancer de la prostate (0.85), Haberman (0.9) et Dermatologie (0.98). Ces résultats aident l'expert dans le processus de diagnostic. À partir de ce tableau, nous remarquons que la méthode proposée atteint de meilleures performances en termes de précision selon les trois algorithmes utilisés.

D'après les taux de classification des 11 bases de données présentés dans la figure 4.4.2, nous observons que l'algorithme CBR-WRF donne de meilleurs résultats que les deux autres algorithmes dans les bases de données suivantes : Diabète (81%), Hépatite (91%), Heart (83%), Maladie du sein (91%) et Cancer de la prostate (85%). Autrement dit, l'utilisation des forêts aléatoires pondérées dans la phase de remémoration augmente la performance du système RàPC.

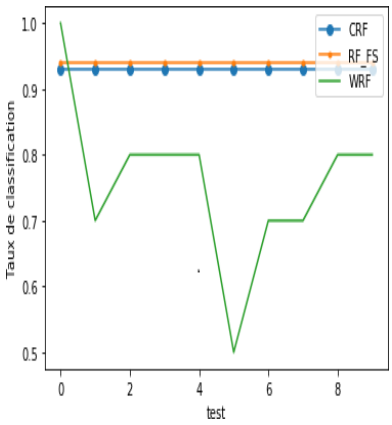
De plus, la précision de l'algorithme CBR-RF-FS montre de bons résultats dans les bases de données suivantes : EEG (94%), Alzheimer (100%), Haberman (82%) et Dermatologie (97%). Cela signifie que l'algorithme CBR-RF-FS améliore également la performance du système RàPC.



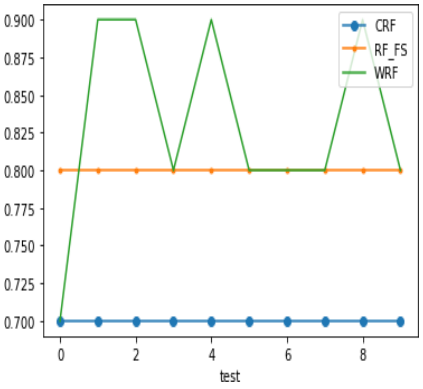
Diabète



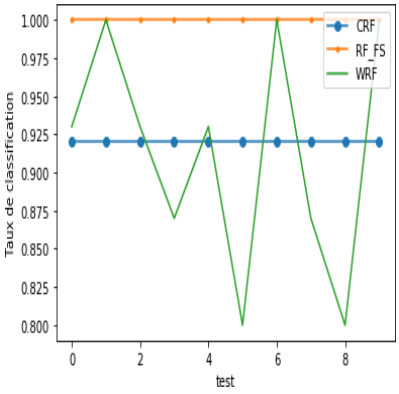
Hépatite



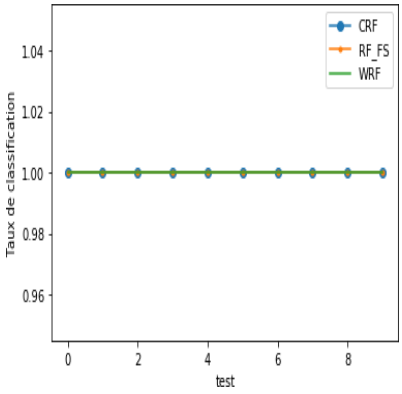
EEG



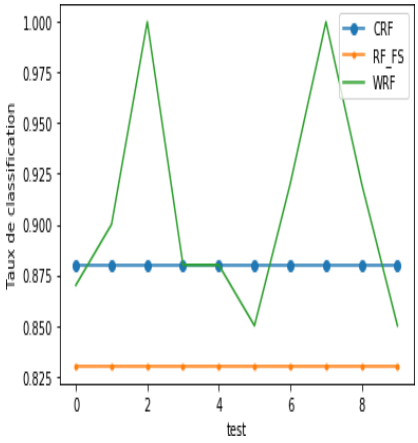
Heart



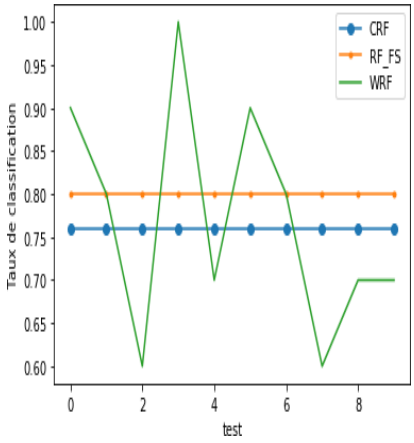
Alzheimer



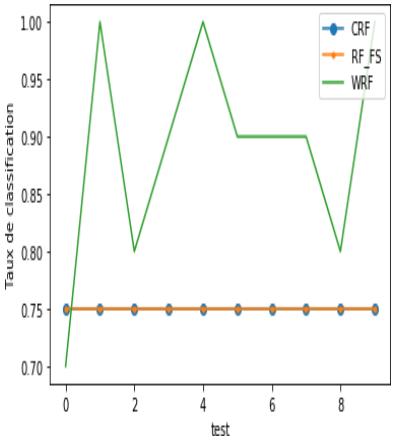
Lung cancer



Maladie de sein



Transfusion du sang



Prostate cancer

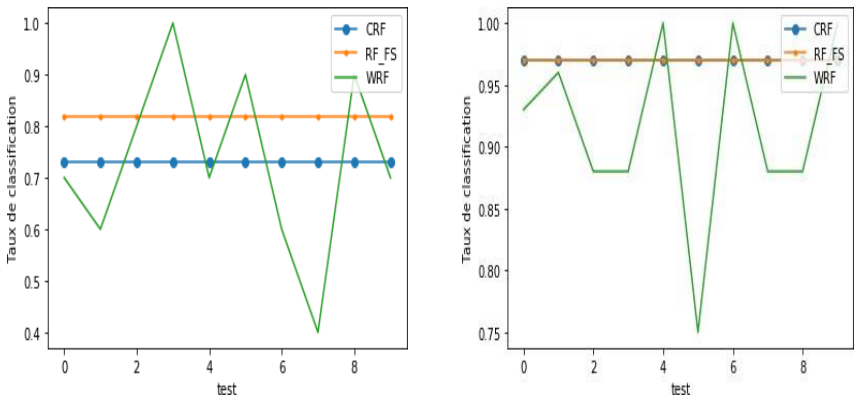


Figure 4.4.2. Courbes d'apprentissage des algorithmes proposés (Tarchoune et al, 2022)

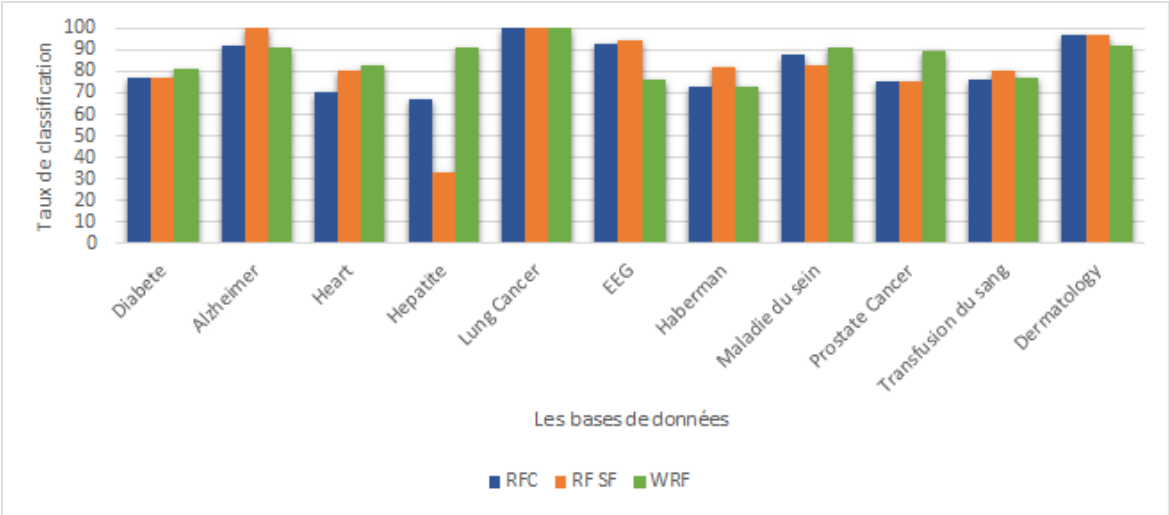


Figure 4.4.3. Histogramme montrant la comparaison entre les taux de classification obtenus pour les onze bases de données

Expérimentation 3

Dans cette expérimentation, notre objectif est d'améliorer la performance par rapport à l'expérimentation précédente. Le tableau 4.4.5 présente les résultats des deux types de forêts (classique et pondérée) en fonction du taux de classification pour différents nombres d'arbres (nbtree=50, 100, 200). Nous observons que la forêt aléatoire pondérée (WRF) améliore la performance du système RàPC, notamment pour les bases de données Pima, Heart et Alzheimer, à différents nombres d'arbres. Les meilleurs résultats ont été obtenus avec des forêts contenant 100 arbres de décision.

Tableau 4.4.5. La performance des algorithmes en changeant le nombre d'arbres (Tarchoune et al, 2022)

La Base de données	nbtree(Nombred'arbres)	TCCBR-CRF	TCCBR-WRF
Pima	50	0.77	0.87
	100	0.77	0.89
	200	0.76	0.89
Hépatite	50	0.9	0.9
	100	0.96	0.9
	200	0.94	0.9
Lung cancer	50	0.9	0.9
	100	0.9	0.9
	200	0.9	0.9
Heart	50	0.82	0.8
	100	0.86	0.9
	200	0.85	0.9
Alzheimer	50	0.77	0.7
	100	0.76	0.8
	200	0.76	0.9
EEG	50	0.92	0.6
	100	0.93	0.8
	200	0.93	0.8
Maladie du sein	50	0.94	0.7
	100	0.94	0.8
	200	0.94	0.8
Transfusion du sang	50	0.76	0.74
	100	0.77	0.75
	200	0.76	0.75
Prostate cancer	50	0.91	0.79
	100	0.94	0.81
	200	0.94	0.82
Dermatology	50	0.47	0.8
	100	0.47	0.8
	200	0.47	0.8
Haberman	50	0.67	0.64
	100	0.62	0.66
	200	0.64	0.66

4.4.6 Résultats et discussions

D'après l'histogramme de la figure 4.4.3, CBR-WRF a été meilleure dans la classification des données de 5 bases:

- Diabète
- Heart
- Hépatite
- Maladie du sein
- Prostate Cancer

CBR-RF-FS a été la meilleure dans la classification de quatre bases de données :

- Alzheimer
- EEG
- Habermanet
- Transfusion de sang

Les trois algorithmes ont montré des performances équivalentes dans la classification de la base « Lung Cancer ». Cependant, CBR-CRF et CBR-RF-FS ont obtenu de meilleurs résultats que CBR-WRF dans la classification de la base « Dermatology », tandis que CBR-CRF n'a surpassé aucun des autres algorithmes dans la classification de quelconque base. En conséquence, les algorithmes CBR-WRF et CBR-RF-FS ont nettement surpassé CBR-CRF, indiquant ainsi une amélioration notable en passant des méthodes classiques à la sélection d'attributs et à la pondération. Notons également que CBR-WRF a surpassé CBR-RF-FS de cinq points, se positionnant ainsi en tête parmi les trois méthodes.

Cette amélioration était attendue, car les méthodes CBR-RF-FS et CBR-WRF mettent en avant les attributs les plus importants, en écartant les moins déterminants ou en attribuant des poids plus élevés aux attributs les plus significatifs pour leur impact dans la classification. Plus précisément, avec CBR-RF-FS, les attributs les moins importants ont été éliminés, ne conservant que les plus pertinents. En revanche, avec CBR-WRF, tous les attributs ont été conservés mais les plus importants ont reçu un poids renforcé, ce qui a été réalisé en multipliant l'entropie par le poids correspondant à chaque attribut. Le poids W est défini comme le nombre d'attributs importants divisé par le nombre total d'attributs.

Nous avons présenté une forêt aléatoire pondérée qui a montré une amélioration significative dans la phase de remémoration, et donc du système RàPC. Dans le tableau 4.4.6, nous comparons les précisions de l'approche proposée avec celles d'autres méthodes basées sur RF. Nous constatons que notre approche a atteint une précision comparable, voire supérieure, à celle des autres méthodes de comparaison dans la majorité des bases de données, prouvant ainsi l'efficacité de l'intégration du classifieur forêt aléatoire avec une performance de classification élevée.

Tableau 4.4.6. Résultats comparatifs de l'approche proposée par rapport à d'autres techniques d'apprentissage (Tarchoune et al, 2022)

Méthodes	**Travaux connexes**	**Ensemble de données**	**Précision (%)**
CBR-RF	CBR, RF (Darabi et al, 2014)	Asthme	80
	CBR, RF (Ayeldeen et al, 2015)	Cancer du sein	99
	CBR, RF (Asim et al, 2019)	Jeu de données standard du blogueur	95
Notre approche	CBR-CRF	11 bases	97
	CBR-RF-RS	11 bases	97
	CBR-WRF	11 bases	**100**

Dans le tableau 4.4.6, nous comparons la précision de l'approche proposée avec celle des différentes méthodes basées sur la forêt aléatoire (RF). Il est évident que Darabi et al, 2014 ont testé leur méthode sur 325 cas asthmatiques et non asthmatiques, obtenant une précision de 80%. Ayeldeen et al, 2015 ont testé leur méthode sur des cas de cancer du sein, atteignant une précision de 99%. De même, Asim et al, 2019 ont obtenu une précision de 95% en utilisant les données des blogueurs.

Lorsque nous avons testé notre contribution sur 11 bases de données, nous avons obtenu une précision de 97% avec le premier algorithme CBR-CRF, 97% avec le deuxième algorithme CBR-RF-FS, et 100% avec le troisième algorithme CBR-WRF. Bien que les travaux des chercheurs précédents aient montré de très bonnes précisions, il est important de noter qu'ils ont testé leurs méthodes sur une seule base de données. En revanche, notre approche a été évaluée sur 11 bases différentes. Si nous avions testé nos algorithmes sur une seule base, nous aurions probablement obtenu des précisions égales ou supérieures à celles rapportées par ces chercheurs.

En comparaison, notre approche a montré une précision comparable, voire meilleure, dans la majorité des bases de données. Cela démontre que l'intégration du classifieur RF dans la phase de remémoration est efficace, offrant une performance de classification élevée.

4.4.7 Conclusion

L’approche proposée vise à modéliser la phase de remémoration du système RàPC en utilisant le classifieur forêt aléatoire (RF), en se basant sur trois algorithmes différents. Le premier est la forêt aléatoire classique, le deuxième est la forêt aléatoire avec sélection des attributs, où les attributs les plus importants sont sélectionnés et les moins importants sont supprimés. Le troisième algorithme est la forêt aléatoire pondérée, où les attributs les plus importants sont pondérés en leur attribuant un poids, ce poids étant multiplié par l'entropie correspondante à chaque attribut.

Nous avons évalué et testé les performances de chaque algorithme sur 11 bases de données médicales différentes, en termes de sensibilité (SE), spécificité (SP), taux de classification (TC) et taux d’erreur (TE). Nous avons constaté que CBR-WRF et CBR-RF-FS donnent de meilleurs résultats que CBR-CRF. Ces résultats confirment que l’intégration du classifieur RF dans la phase de remémoration du système RàPC a permis d’améliorer significativement la performance du système.

4.5 Contribution 4 : Une nouvelle forêt aléatoire améliorée pour la classification des données médicales utilisant la corrélation de Pearson et le meilleur nombre d'arbres

Dans cette contribution, nous avons proposé trois solutions pour la prédiction des données médicales. La première solution optimise un modèle de forêt aléatoire en utilisant une mesure de similarité. La deuxième solution améliore la forêt aléatoire grâce à une utilisation conjointe de la sélection des caractéristiques. Enfin, la troisième approche combine la sélection des caractéristiques avec des mesures de similarité, en se basant sur les forêts aléatoires. Nous démontrons que les performances de prédiction et les taux de classification des forêts aléatoires, lorsqu'elles sont appliquées à onze bases de données extraites de Kaggle et UCI, peuvent être significativement améliorés par ces méthodes d'apprentissage.

4.5.1 Architecture générale de l'approche proposée

Comme le montre la figure 4.5.1, la méthodologie suivie dans cette contribution s'est déroulée en trois étapes principales (1) le prétraitement des données (2) le développement du modèle (3) l'évaluation des résultats.

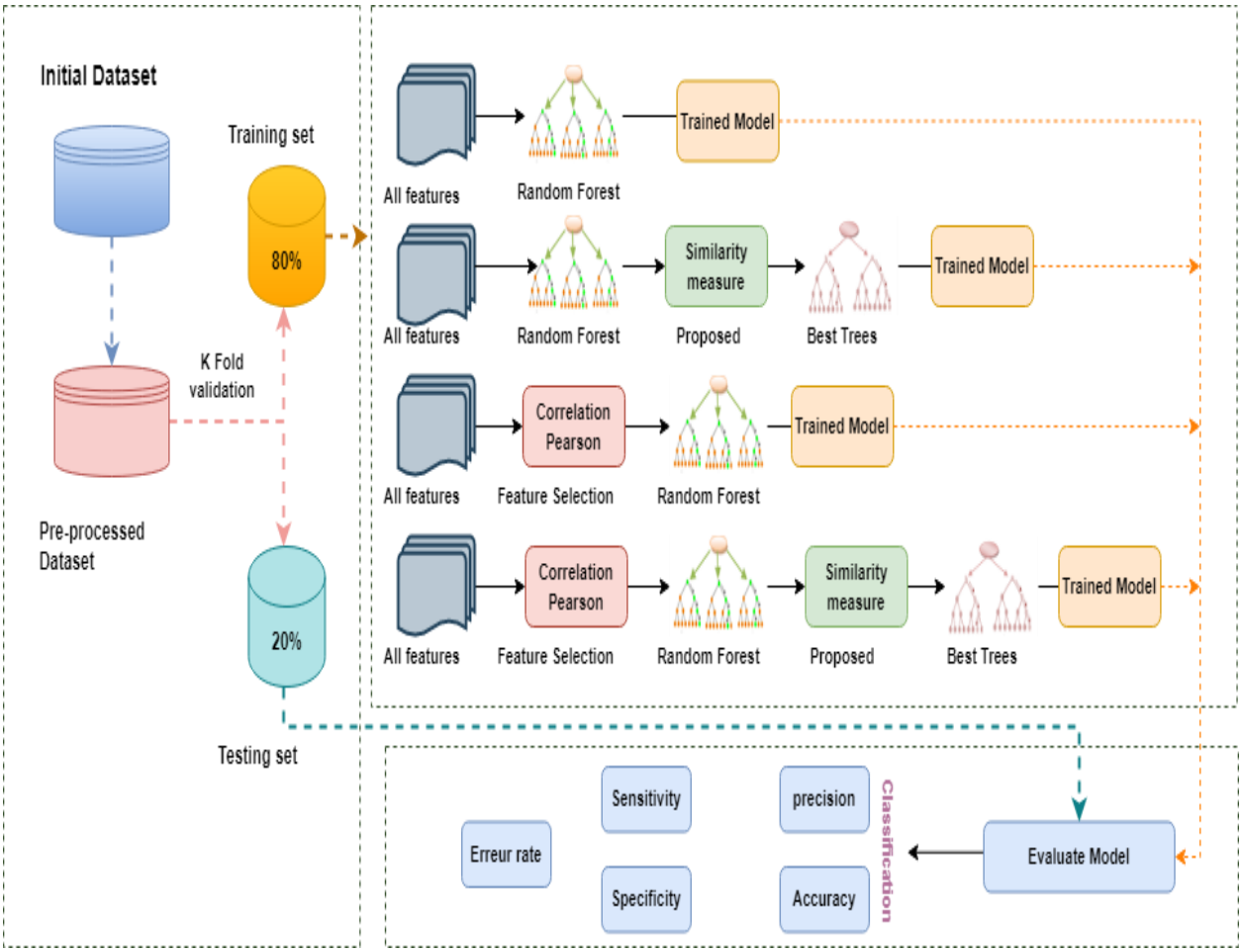


Figure 4.5.1. Architecture générale de l'approche proposée (Tarchoune et al, 2024)

4.5.2 Ensemble de données

Les ensembles de données utilisés dans l'expérience ont été détaillé dans le tableau (tableau 4.5.1), nous avons utilisé 11 bases de données médicales collectés à partir de dépôt d'apprentissage automatique de l'université de Californie à Irvine (UCI), de Kaggle, et une base de l'Algérie.

Tableau 4.5.1. Description des bases de données utilisées

Base de données	Taille	N°Attribut	Réference
Heart	303	14	UCI
Lung cancer	59	7	Kaggle
Hepatitis	155	20	UCI
Breast Cancer	100	28	Algeria
Alzheimer	354	15	UCI
EEG-EYE-STATE	1498	15	Kaggle
Transfusion	748	5	Kaggle
Dermatology	366	35	UCI
Prostate Cancer	100	10	UCI
Haberman	306	4	UCI
Diabetes	768	9	UCI

4.5.3 Résultats et discussions

Dans cette section, nous évaluons la classification multi-classe des bases médicales en utilisant la méthode des forêts aléatoires. Nous comparons les résultats obtenus par différentes variantes : RF_Standard, RF_Similarity, RF_FS et RF_FS_Similarity pour chaque base de données.

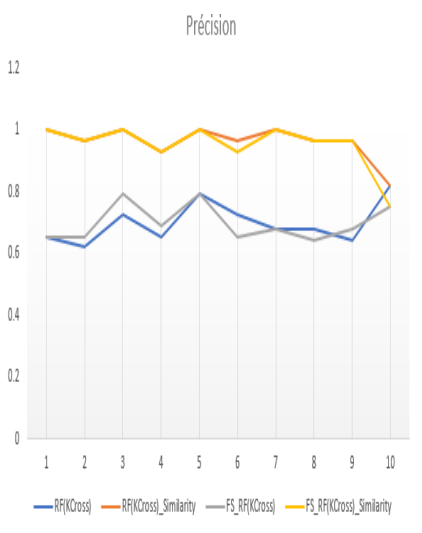
Pour nos expérimentations, nous avons utilisé une forêt aléatoire classique ainsi que trois variantes améliorées. Les onze bases de données utilisées pour les tests sont les suivantes : Breast Cancer, Dermatology, EEG_EYE_STATE, Haberman, Diabetes, Heart, Hepatitis, Lung Cancer, Alzheimer, Prostate Cancer, et Transfusion. Pour chaque forêt, nous avons utilisé 100 arbres de décision (nbtree).

Nous avons construit quatre types de forêts aléatoires différentes avec 100 arbres chacune. La première forêt est une forêt aléatoire classique, utilisée avec la technique de validation croisée en 10-fold pour diviser les données et effectuer 10 tests différents. La deuxième forêt intègre une mesure de similarité. La troisième améliore la méthode classique en ajoutant une sélection des attributs. La quatrième combinaison ajoute une mesure de similarité à la méthode de sélection des attributs.

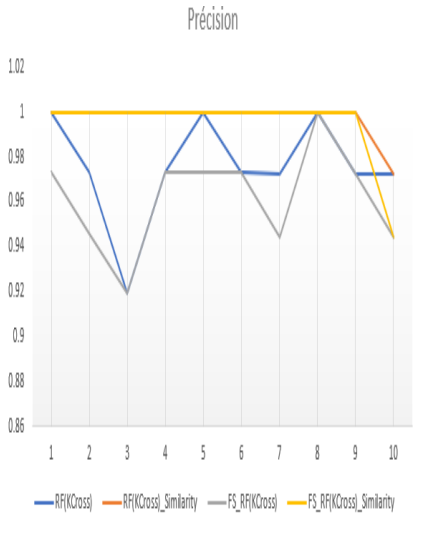
Nous avons exécuté le programme dix fois pour chaque type de forêt afin de comparer les performances des quatre méthodes sur les onze bases de données. Les résultats montrent que les forêts améliorées fournissent de meilleurs résultats que la méthode classique pour toutes les bases de données.

Tableau 4.5.2. Comparaisons des précisions obtenues avec les quatre méthodes utilisées

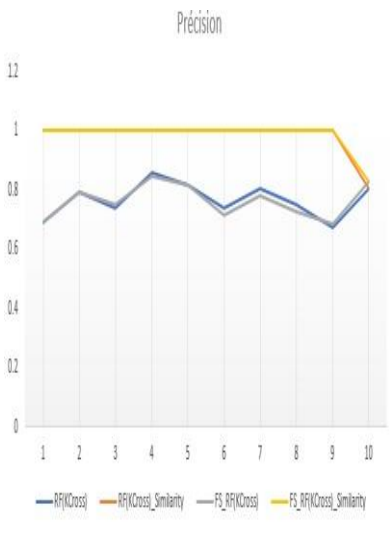
Méthodes Base de données	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast Cancer	69,94	96,12	69,91	95,06
Dermatology	97,55	99,72	96,18	99,44
Diabetes	76,68	98,03	76,30	98,29
EEG-Eye-State	93,20	99,25	88,84	98,88
Haberman	68,25	95,03	68,58	94,69
Heart	33,00	94,00	33,00	94,00
Hepatitis	81,29	96,67	82,58	96,67
Lung cancer	95,00	100	93,33	100
Alzheimer	91,46	99,73	90,65	99,46
Prostate cancer	83,00	99,00	82,00	98,00
Transfusion	74,20	91,16	74,47	91,16



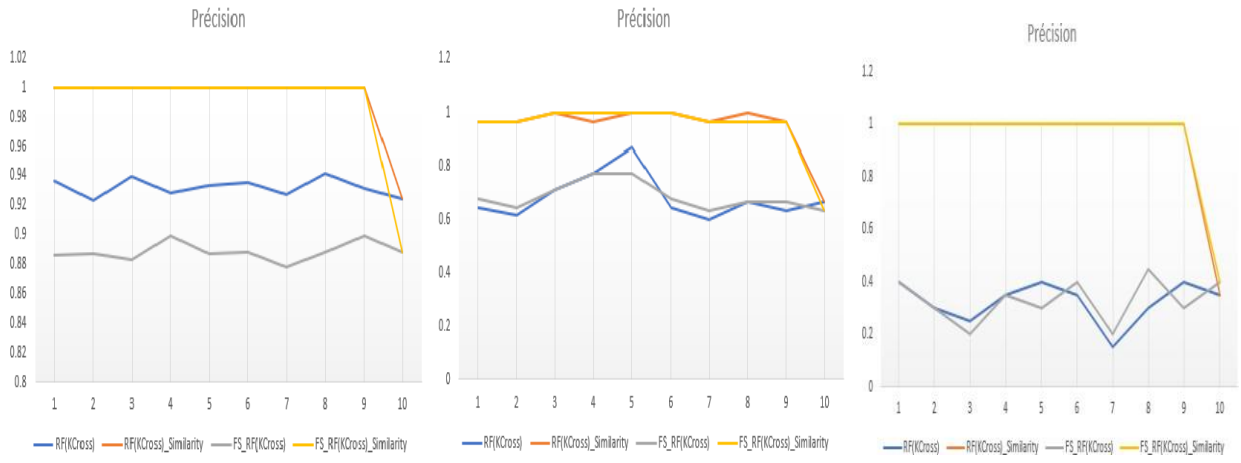
Breast Cancer



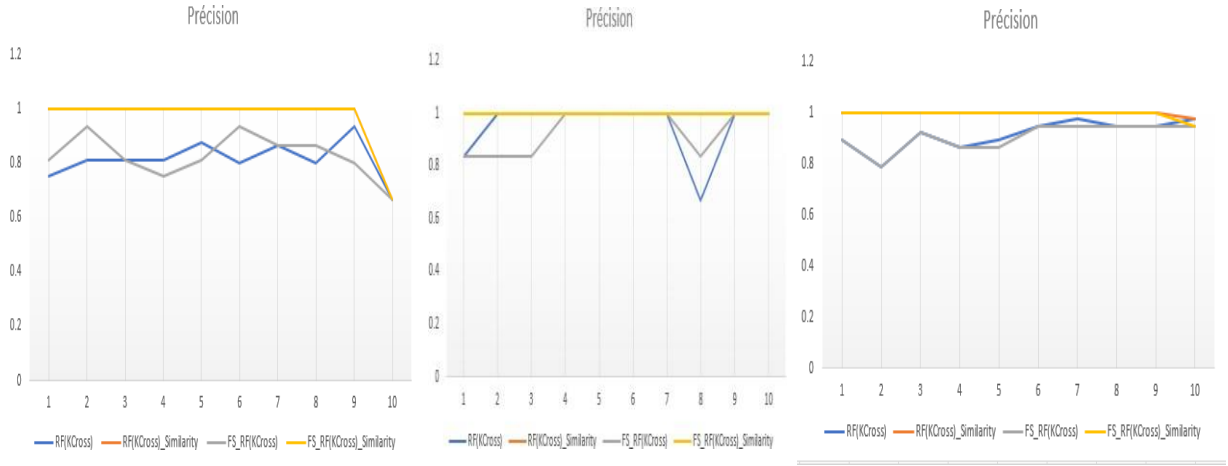
Dermatology



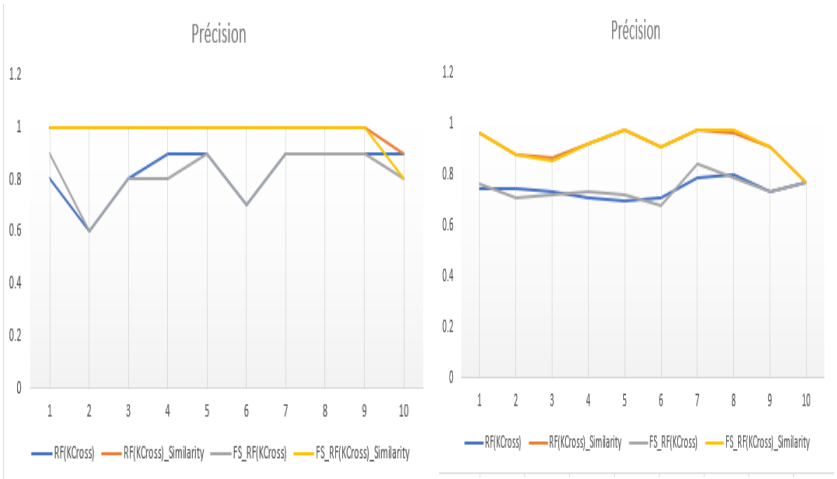
Diabetes



EEG-Eye-State **Haberman** **Heart**



Hepatitis **Lung cancer** **Alzheimer**



Prostate cancer **Transfusion**

Figure 4.5.2. Précision obtenue par les quatre méthodes de classification sur les 11 bases médicales

Dans le tableau 4.5.2 et la figure 4.5.2, nous comparons les performances de précision des quatre variantes de forêts aléatoires. Les résultats montrent les taux de précision pour chaque critère d'évaluation. Nous observons une stabilisation des taux de précision pour la forêt aléatoire classique ainsi que pour la méthode de sélection des attributs sans mesure de similarité, lorsque 100 arbres de décision sont utilisés. En revanche, les deux variantes de classification intégrant la mesure de similarité présentent généralement de meilleurs résultats par rapport aux deux autres méthodes.

Tableau 4.5.3. Résultats des taux de classification obtenus par les quatre méthodes de classification sur 11 bases de données médicales

Méthodes **Base de données**	**RF_Standard**	**RF_Similarity**	**RF_FS**	**RF_FS_Similarity**
Breast Cancer	69,94	96,12	69,91	95,06
Dermatology	36,11	36,11	38,89	38,89
Diabetes	76,68	98,03	76,30	98,29
EEG-Eye-State	93,20	99,25	88,84	98,88
Haberman	68,25	95,03	68,58	94,69
Heart	20,00	51,00	20,00	51,00
Hepatitis	81,29	96,67	82,58	96,67
Lung cancer	95,00	100	93,33	100
Alzheimer	41,02	48,76	40,21	48,49
Prostate cancer	83,00	99,00	82,00	98,00
Transfusion	74,20	91,16	74,46	91,16

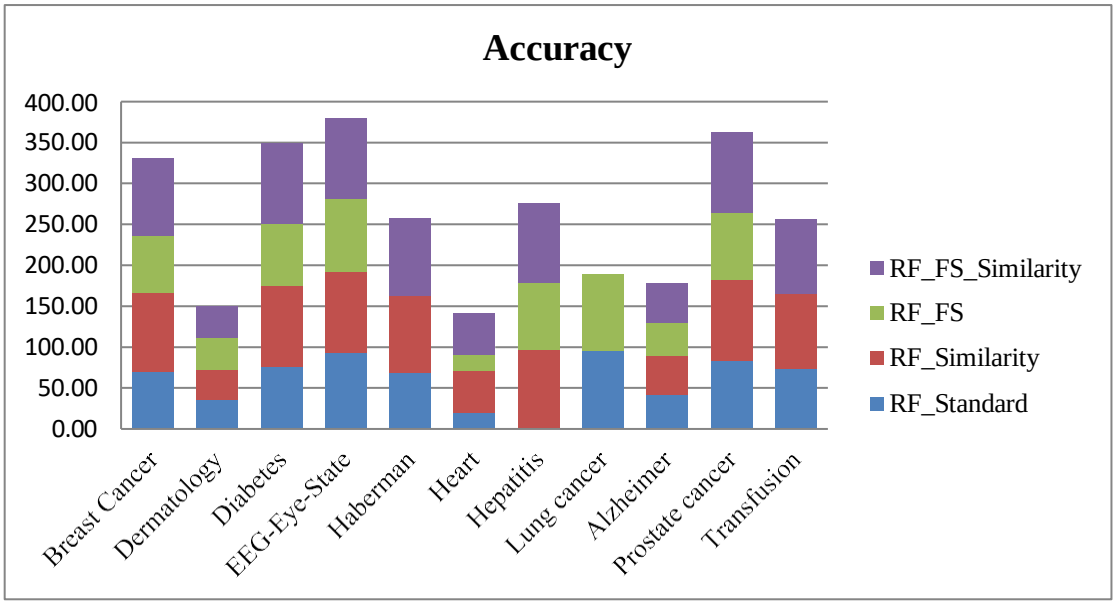


Figure 4.5.3. Histogramme comparatif des résultats de taux de classification obtenu par les quatre méthodes de classification sur les bases de données utilisées

Dans le tableau 4.5.3 et la figure 4.5.3, nous comparons les performances de taux de classification des quatre différentes forêts aléatoires. Nous présentons les résultats des taux de classification pour chaque critère d'évaluation, nous remarquons une stabilisation des taux de classification pour 100 arbres

de décisions de la méthode classique et la méthode de sélections des attributs sans mesure de similarité. Nous remarquons également que les deux classifications avec la mesure de similarité fournissent la plupart du temps de meilleurs résultats par rapport aux deux autres.

Tableau 4.5.4. Résultats des sensitivités obtenus par les quatre méthodes de classification sur 11 bases de données médicales

Méthodes Base de données	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast Cancer	34,28	92,33	48,73	92,33
Dermatology	98	98,00	100	100
Diabetes	60,77	96,67	69,13	98,18
EEG-Eye-State	89,84	98,78	89,80	99,11
Haberman	23,34	88,02	34,84	90,40
Heart	56,00	96,00	58,50	100
Hepatitis	91,77	98,18	85,46	98,00
Lung cancer	85,00	100	84,17	100
Alzheimer	99,33	100	90,45	99,33
Prostate cancer	86,29	100	87,40	96,67
Transfusion	28,66	69,92	42,48	88,66

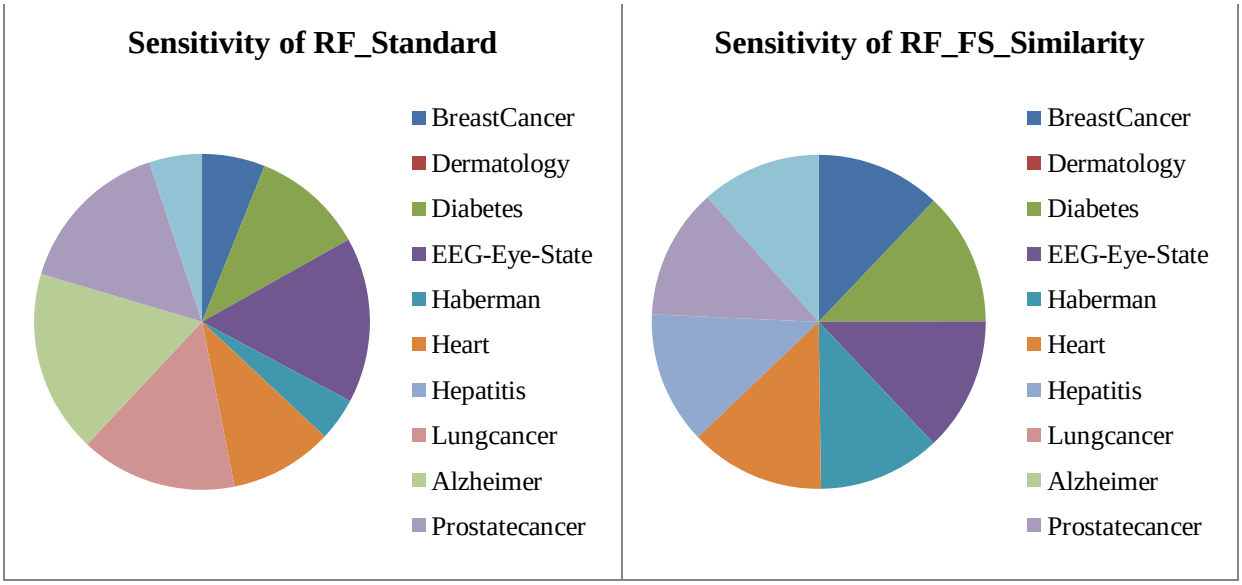


Figure 4.5.4. Histogramme comparatif des résultats de la sensitivité obtenue par les deux méthodes de classification (Standard RF et Improved RF) sur les bases de données utilisées

Tableau 4.5.5. Résultats de spécificités obtenues par les quatre méthodes de classification sur 11 bases de données médicales

Méthodes Base de données	RF_Standard	RF_Similarity	RF_FS	RF_FS_Similarity
Breast Cancer	85,57	97,64	75,56	96,49
Dermatology	100	100	100	100
Diabetes	85,15	98,78	79,67	98,33
EEG-Eye-State	95,94	99,63	88,11	98,71
Haberman	84,36	97,00	75,68	96,83
Heart	55,19	95,00	45,83	95,00
Hepatitis	43,17	93,00	71,67	93,33
Lung cancer	98,00	100	96,33	100
Alzheimer	33,33	100	30	9,00
Prostate cancer	70	98,33	65,24	100
Transfusion	88,13	97,90	80,28	91,56

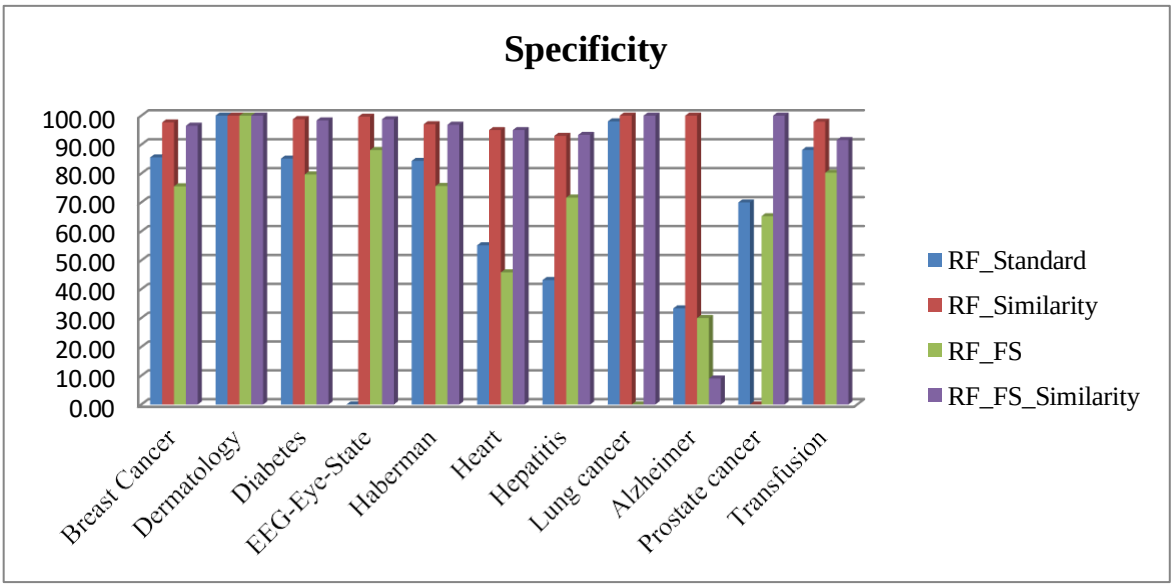


Figure 4.5.5**.** Histogramme comparatif des résultats de la spécificité obtenue par les quatre méthodes de classification sur 11 bases de données

Dans cette contribution, nous avons développé et testé quatre approches pour la classification en utilisant des forêts aléatoires :

1. Forêt Aléatoire Classique avec Validation Croisée : La première phase consiste en l'utilisation de la forêt aléatoire classique combinée à la technique de validation croisée en 10-fold.
2. Forêt Aléatoire avec Mesure de Similarité : La deuxième phase améliore le système en utilisant une mesure de similarité, où seuls les arbres de décision ayant des taux de classification élevés sont retenus. Cette méthode permet de renforcer la performance de prédiction par rapport à la méthode classique.
3. Forêt Aléatoire avec Sélection des Attributs : La troisième méthode utilise la technique de sélection des caractéristiques par Corrélation de Pearson. Cette approche sélectionne les caractéristiques importantes tout en éliminant celles non pertinentes, ce qui améliore la vitesse d'entraînement en réduisant la dimensionnalité et en résolvant le problème de haute dimensionnalité. Elle est applicable à différents types de caractéristiques de données. .

4. Forêt Aléatoire avec Sélection des Attributs et Mesure de Similarité : Enfin, la quatrième méthode combine la sélection des attributs avec la mesure de similarité pour une amélioration supplémentaire. Nous avons mené une étude comparative entre ces méthodes en utilisant les mêmes critères de performance pour choisir la meilleure approche.

Les expérimentations montrent que les quatre méthodes ont produit de bons résultats, chacune offrant des améliorations spécifiques en fonction des critères évalués.

Tableau 4.5.6. Résultats des taux d'erreurs obtenus par les quatre méthodes de classification sur 11 bases de données médicales

Méthodes **Base de données**	**RF_Standard**	**RF_Similarity**	**RF_FS**	**RF_FS_Similarity**
Breast Cancer	0.30	0.03	0.30	0.04
Dermatology	0.02	0.02	0	0
Diabetes	0.19	0.19	0.17	0.17
EEG-Eye-State	0.06	0.007	0.11	0.01
Haberman	0.31	0.04	0.31	0.05
Heart	0.17	0.02	0.19	0.01
Hepatitis	0.33	0.33	0.33	0.33
Lung cancer	0.05	0.0	0.06	0.0
Alzheimer	0.03	0.0	0.04	0.002
Prostate cancer	0.17	0.01	0.18	0.02
Transfusion	0.25	0.08	0.25	0.08

Les tableaux 4.5.4 et 4.5.5 présentent les sensitivités et spécificités obtenues pour chaque algorithme testé. En comparant les méthodes proposées, nous constatons que les deux méthodes intégrant la mesure de similarité ont atteint des performances supérieures, avec des taux de précision allant jusqu'à 100%.

Le tableau 4.5.6 montre les taux d'erreur pour chaque méthode sur toutes les bases de données. Nous remarquons que les deux méthodes utilisant la mesure de similarité obtiennent les meilleurs taux d'erreur. En particulier, la méthode RF_FS_Similarity atteint un taux d'erreur de 0 sur les bases Dermatology et Lung Cancer, tandis que la méthode RF_Similarity obtient également un taux d'erreur de 0 pour les bases Lung Cancer et Alzheimer.

Tableau 4.5.7. Les résultats comparatifs de l'approche proposée contre les autres techniques d'apprentissage

Travaux connexes	**Techniques utilisées**	**Bases utilisées**	**Précision**
Rawashdeh et al, 2020	RF classique	Naissance prématurée	86%
Liu et al, 2020	Correlation + RF classique	Rythme cardiaque fœtal	97%
Li et al, 2020	Correlation + RF classique	Phalogramme d'électroenc Généralisé postictal	94%
Bi et al, 2020	Pearson + RF classique	Alzheimer	86%
Notre approche			
RF_Standard	RF classique	11 bases médicales	95%
RF_Similarity	RF classique + Similarity Mesure		**100%**
RF_FS	RF classique + Pearson Correlation		93%
RF_FS_Similarity	RF classique + Pearson Correlation + Similarity Mesure		**100%**

La forêt aléatoire est un modèle d'apprentissage intégré reconnu pour ses avantages, bien que de nombreux travaux continuent de chercher à l'améliorer. Par exemple, Rawashedeh et al, 2020 ont appliqué un modèle de forêt aléatoire classique, mais n'ont atteint qu'un taux de classification de 86%. Liu et al, 2020 et Li et al, 2020 ont utilisé des forêts aléatoires classiques combinées à une technique de sélection de caractéristiques (Corrélation), obtenant des taux de classification de 97% et 94% respectivement. Bi et al, 2020 ont appliqué une forêt aléatoire améliorée avec la technique Pearson sur une seule base de données, atteignant également 86% de précision.
En revanche, le modèle de forêt aléatoire proposé dans cette contribution, avec ses trois variantes, présente un effet prédictif supérieur sur 11 bases de données médicales. Les détails des techniques utilisées et leurs performances comparées à d'autres études sont présentés dans le tableau 4.5.7.
Cette analyse comparative montre que notre modèle proposé est compétitif par rapport aux différents modèles de classification présents dans la littérature. De plus, l'implémentation de la forêt aléatoire améliorée dans un environnement clinique pourrait assister les médecins dans leurs prises de décisions cliniques.

4.5.4 Conclusion

Dans cette contribution, nous avons étudié la performance d'un modèle ensembliste appelé forêt aléatoire dans le contexte de la classification des données médicales. Notre analyse des travaux existants a mis en évidence les avantages et les limites des forêts aléatoires utilisées pour la classification dans ce domaine.

Nous avons constaté que, bien que les classifieurs basés sur les forêts aléatoires montrent de bonnes performances, il est possible de les améliorer pour obtenir des résultats encore plus précis. Pour cela, nous avons proposé trois méthodes utilisant différentes variantes des forêts aléatoires. Nous avons évalué notre modèle sur onze bases de données provenant d'UCI et Kaggle.

Nous avons d'abord ré-implémenté le modèle de forêt aléatoire classique en utilisant la technique de K-Cross validation. Ensuite, nous avons développé plusieurs variantes de ce classifieur en intégrant des méthodes de sélection des attributs et une mesure de similarité.

Nous avons évalué et testé les performances de chaque variante en termes de précision, sensibilité, spécificité, taux de classification et taux d'erreur. Les résultats obtenus avec nos quatre méthodes sont parmi les meilleurs pour la classification de ces bases de données et se montrent très compétitifs par rapport aux autres versions des forêts aléatoires.

Conclusion Générale & Perspectives

Conclusion générale et perspectives

Dans cette thèse, nous avons exploré le domaine du raisonnement à partir de cas (RàPC), en nous concentrant sur trois connecteurs de connaissance essentiels : la base de cas, les mesures de similarité, et les règles d'adaptation. Nous avons intégré les forêts aléatoires dans l'approche RàPC appliquée au diagnostic médical, en proposant une méthode de remémoration guidée par l'adaptation du système RàPC basée sur les forêts aléatoires.

Nous avons d'abord réalisé un état de l'art sur les travaux existants concernant le raisonnement à partir de cas, les forêts aléatoires, et les techniques de sélection de caractéristiques, avec une attention particulière aux applications médicales.

L'objectif de notre travail est de modéliser les connaissances pour soutenir le diagnostic de diverses maladies. Nous avons structuré notre approche en deux contributions principales :

1. Première Contribution : Nous avons modélisé la base de cas et la phase de remémoration du raisonnement à partir de cas en utilisant des forêts aléatoires améliorées. Cette contribution implique la modélisation des cas à l'aide de techniques de sélection de caractéristiques manuelles ou de pondération des caractéristiques, suivie de la modélisation de la remémoration avec des forêts aléatoires optimisées. Cette approche permet de mieux gérer la complexité du modèle, rendant ainsi la remémoration plus efficace.
2. Deuxième Contribution : Nous avons proposé un algorithme d'adaptation basé sur des règles et une mesure d'adaptation, destiné à ajuster le cas remémoré au cas actuel.

Les résultats expérimentaux montrent que l'intégration des techniques automatiques de sélection des caractéristiques est prometteuse, car elle réduit la taille de la base de cas et le temps de remémoration, tout en maintenant un niveau élevé de précision de classification.

L'intégration des forêts aléatoires améliorées dans le raisonnement à partir de cas répond efficacement aux besoins des cliniciens en comblant les lacunes des systèmes existants. Toutefois, plusieurs perspectives de développement restent ouvertes :

- Validation avec d'autres Méthodes : Il serait intéressant de valider cette méthode en explorant d'autres combinaisons de méthodes pour la décision finale, telles que le vote pondéré ou le vote moyen, afin d'améliorer l'exactitude et la performance des systèmes de classification.
- Techniques Alternatives de Sélection de Caractéristiques : L'utilisation de techniques alternatives pour la sélection des caractéristiques pourrait permettre de révéler de nouvelles relations entre les fonctionnalités.
- Développement des Autres Phases du Système RàPC : Il serait pertinent d'étendre le développement aux autres phases du système de raisonnement à partir de cas, telles que la révision et le maintien, pour renforcer l'approche proposée.

Références bibliographiques

(Aamodt et Plaza, 1994) Aamodt, A., & Plaza, E. (1994).Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. http://content.iospress.com/articles/ai-communications/aic7-1-04.

(Admass et Munaye, 2024) Admass, W.S., Munaye, Y.Y. Integrating case-based and rule-based reasoning for diagnosis and treatment of mango disease using data mining techniques. Int. j. inf. tecnol. 16, 1699–1715 (2024). https://doi.org/10.1007/s41870-023-01587-y.

(Akin Ozcift, 2011) Ozcift, A. (2012). «Enhanced Cancer Recognition System Based on Random Forests Feature Elimination Algorithm», J. Med. Syst., vol. 36, no 4, p. 2577‑2585, doi: 10.1007/s10916-011-9730-1.

(Alsun, 2012) Al Sun, M. H. A. (2012). Indexation guidée par les connaissances enimagerie médicale [Phd thesis, Télécom Bretagne, Université de Bretagne Occidentale]. https://tel.archives-ouvertes.fr/tel-00719587

(Alam et al, 2020) Alam, Md. Z., Rahman, M. S., & Rahman, M. S. (2019). «A Random Forest based predictor for medical data classification using feature ranking», Inform. Med. Unlocked, vol. 15, p. 100180, doi: 10.1016/j.imu.2019.100180.

(Althoff, 2011) AlthoffKlaus-Dieter.2011. «CASE-BASED-REASONING».

(Anirudh et al, 2019) Hebbar, AP, Kumar M, M V, & SanjayH A (2019). «DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes», in1st International Conference on Advances inInformation Technology (ICAIT), Chikmagalur, India,p. 271‑276, doi: 10.1109/ICAIT47043.2019.8987277.

(Asadi et al, 2021) Asadi, S., Roshan, S., & Kattan, M. W. (2021). «Random forest swarm optimization-based for heart diseases diagnosis», J. Biomed.Inform., vol. 115, p. 103690, doi: 10.1016/j.jbi.2021.103690.

(Asim et al, 2019) Asim,Y., Raza,B. Malik,A. K., Shahaid,A. R.,& Alquhayz,H.(2019). «An Adaptive Model for Identification of Influential Bloggers Based on Case-Based Reasoning Using Random Forest », IEEE Access, vol. 7, p. 87732‑87749, doi: 10.1109/ACCESS.2019.2925905.

(Ayeldeen et al, 2015) Ayeldeen, H., Elfattah, M. A., Shaker, O., Hassanien, A. E., & Kim, T.-H. (2015). «Case-Based Retrieval Approach of Clinical Breast Cancer Patients», in 2015 3rd International Conference on Computer, Information and Application, Yeosu, South Korea, p. 38-41.doi: 10.1109/CIA.2015.17.

(Bach et Mork, 2021) Bach Kerstin, et Paul Jarle Mork. s. d. « On the Explanation of Similarity for Developing and Deploying CBR Systems ».

(Bach et Mork, 2021) Bach Kerstin, et Paul Jarle Mork. s. d. « On the Explanation of Similarity for Developing and Deploying CBR Systems ».

(Begum et al, 2011) Begum, S. & Mälardalenshögskola.(2011).A personalized case-based stress diagnosis system using physiological sensor signals.School of Innovation, Design and Engineering, Mälardalen University. http://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-12257.

(Benamina et al, 2018) Benamina, Mohammed, Baghdad Atmani, et Sofia Benbelkacem. 2018. « Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic ». International Journal of Interactive Multimedia and Artificial Intelligence 5 (3): 72. https://doi.org/10.9781/ijimai.2018.02.001.

(Benbelkacem et Atmani, 2019) Benbelkacem, S., & Atmani, B. (2019). «Random Forestsfor Diabetes Diagnosis», in 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, p. 1‑4, doi: 10.1109/ICCISci.2019.8716405.

(Bentaiba et al, 2018) Bentaiba-Lagrid, Miled Basma, Lydia Bouzar-Benlabiod, Stuart H. Rubin, Thouraya Bouabana- Tebibel, et Maria Roumaissa Hanini. 2018. « Knowledge Amplification Using Randomization in Case-Based Reasoning: Case Study on Severity of Mammography Mass. ». In 2018 IEEE International Conference on Information Reuse and Integration (IRI), 155‑62. Salt Lake City, UT: IEEE. https://doi.org/10.1109/IRI.2018.0003.

(Bentaiba et al, 2020) Bentaiba-Lagrid, Miled Basma, Lydia Bouzar-Benlabiod, Stuart H. Rubin, Thouraya Bouabana- Tebibel, et Maria R. Hanini. 2020. « A Case-Based Reasoning System for Supervised Classification Problems in the Medical Field ». Expert Systems with Applications 150 (juillet): 113335. https://doi.org/10.1016/j.eswa.2020.113335.

(Bhalaji et al, 2018) Bhalaji, N., Kumar, K.B.S., Selvaraj, C. (2018). Empirical study of feature selection methods over classificationalgorithms, Int. J. Intelligent Systems Technologies And Applications, Vol. 17, Nos. 1-2, https://doi.org/10.1504/IJISTA.2018.091590.

(Bichindaritz et Montani, 2011) Bichindaritz, Isabelle, et Stefania Montani. 2011. « Advances inCase-Based Reasoning in the Health Sciences ». Artificial Intelligence in Medicine 51 (2): 75‑79. https://doi.org/10.1016/j.artmed.2011.01.001.

(Blanco et al, 2013) Blanco, Xiomara, Sara Rodríguez, Juan M. Corchado, et Carolina Zato. 2013. « Case-Based Reasoning Applied to Medical Diagnosis and Treatment ». In Distributed Computing and Artificial Intelligence, édité par Sigeru Omatu, José Neves, Juan M. Corchado Rodriguez, Juan F Paz Santana, et Sara Rodríguez Gonzalez, 217:137‑46. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00551-5_17

(Bose et al, 2017) S. Bose, V. Rama, N. Warangal, et C. B. Rama Rao, «EEG signalanalysis for Seizuredetectionusing DiscreteWavelet TransformandRandomForest », in2017 International Conference on Computer and Applications (ICCA), Doha, United Arab Emirates, sept. 2017, p. 369‑378.doi: 10.1109/COMAPP.2017.8079760.

(Bouaguel, 2015) Bouaguel, W. (2015).A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data. The 19th Asia Pacific Symposium, Bangkok, Thailand, 75-83.

(Breiman et al, 1984) Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees.Biometrics, 40(3), 874. https://doi.org/10.2307/2530946.

(Breiman, 2001) Breiman, L., (2001). Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

(Chakraborty et al, 2015) Chakraborty, Souvik, Chiranjit Pal, Shambo Chatterjee, Baisakhi Chakraborty, et Nabin Ghoshal. 2015. « Knowledge-Based System Architecture on CBR for Detection of Cholera Disease ». In Intelligent Computing and Applications, édité par Durbadal Mandal, Rajib Kar, Swagatam Das, et Bijaya Ketan Panigrahi, 343:155‑65. Advances in Intelligent Systems and Computing. New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2268-2_17.

(Chattopadhyay et al, 2013) Chattopadhyay, Subhagata, Suvendu Banerjee, Fethi A. Rabhi, et U. Rajendra Acharya. 2013. « A Case-Based Reasoning System for Complex Medical Diagnosis ». Expert Systems 30 (1): 12‑20. https://doi.org/10.1111/j.1468-0394.2012.00618.x.

(Cherry et al, 2014) Cherry, K. M., Wang, S. Turkbey, E. B., & Summers, R. M. (2014). «Abdominal lymphadenopathy detection using random forest», San Diego, California, USA, p. 90351G, doi: 10.1117/12.2043837.

(Choudhury et Begum, 2017) Choudhury, Nabanita, et Dr Shahin Ara Begum. 2017. « THE ROLE OF FUZZY LOGIC IN CASE- BASED REASONING: A SURVEY » 8 (3): 8.

(Darabi et al, 2014) Darabi, S. A. (2014). Case-Based-Reasoning System for Feature Selection and Diagnosing Disease; Case Study: Asthma », p. 18.

(Dash et Liu, 1997) Dash M and LiuH, (1997)."Feature selection for classification," Intelligent Data Analysis, vol. 1, no.3, pp.131–156.

(David, 2003) Leake, D. B. (2003). Case-based reasoning. John Wiley and Sons Ltd. http://dl.acm.org/citation.cfm?id=1074199.

(Deetchakraborty, 2021) De, Sumana, et Baisakhi Chakraborty. 2021. « Case-Based Reasoning (CBR)-Based Anemia Severity Detection System (ASDS) Using Machine Learning Algorithm ». In Advanced Machine Learning Technologies and Applications, 1141:621‑32. Intelligent Systems and Computing.Springer. https://doi.org/10.1007/978-981-15-3383-9_56.

(Devi et al, 2018) Devi, B., Kumar, Anuradha, S., & Shankar, V. G. (2019). «AnaData: A Novel Approach for Data Analytics Using Random Forest Tree and SVM», in Computing, Communication and Signal Processing, vol. 810, B. Iyer, S. L. Nalbalwar, et N. P. Pathak, Éd. Singapore: Springer Singapore, p. 511521.

(Divya et al, 2019) Divya, S., Vignesh, R., Revathy, R. (2019). A Distincitve Model to Classify Tumor Using Random Forest Classifier, IEEE Third International Conference on Inventive Systems and Control(ICISC), pp. 44–47. https://doi.org/10.1109/ICISC44355.2019.9036473.

(Djellali et al, 2018) Djellali, H., Djebbar, A., Zine, N.G., Azizi, N. (2018). Hybrid Artificial Bees Colony and Particle Swarm on Feature Selection, Computational Intelligence and Its Applications, pp. 93–105, https://doi.org/10.1007/978-3-319-89743-1_9.

(Esposito et al, 1993) Esposito, F., Malerba, D., Semeraro, G. (1993).Decision tree pruning as a search in the state space, Machine Learning: ECML-93, pp.165–184.https://doi.org/10.1007/3-540-56602-3_135.

(Farzana et Nooraini, 2013) Ahmad, F. K., & Yusoff, N. (2013). «Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier», in 2013 13th International Conference on Intellient Systems DesignandApplications, Salangor, Malaysia,p. 121,125, doi:10.1109/ISDA.2013.6920720.

(Feuillâtre et al, 2020) Feuillâtre, H., Auffret, V., Castro, M., Lalys, F., Breton, H. L., Garreau, M., & Haigron, P. (2020). Similarities measures and attribute selection for case-based reasoning in transcatheter aortic valve implantation. PLOS ONE, 15(9), e0238463. https://doi.org/10.1371/journal.pone.0238463.

(Gangwar et al, 2024) Gangwar, A.K., Kamthan, P. (2024). Improved Random Forest Classifier for Predicting Heart Disease. In: Santosh, K.C., Sood, S.K., Pandey, H.M., Virmani, C. (eds) Advances in Artificial-Business Analytics and Quantum Machine Learning. COMITCON 2023. Lecture Notes in Electrical Engineering, vol 1191. Springer, Singapore. https://doi.org/10.1007/978-981-97-2508-3_36.

(Girard, 2007) Girard,A.(2007).Exploration d'unalgorithmegénétiqueetd'unarbredecisionàdesfinsde catégorisation[Masters,UniversitéduQuébecàTrois-Rivières].https://depot-e.uqtr.ca/id/eprint/1468/ .

(Golobardes et al, 2002) Golobardes, Elisabet, Xavier Llorà, Maria Salamó, et Joan Martí. 2002. « Computer Aided Diagnosis with Case-Based Reasoning and Genetic Algorithms ». Knowledge-Based Systems 15 (1‑2): 45‑52. https://doi.org/10.1016/S0950-7051(01)00120-4.

(Gu et al, 2020) Guet al.-2020-«Acase-based ensemble learning system for explaina. ».

(Hall et al, 1999) Hall, M. A. (1999). Correlation-based feature selection for machine learning [Thesis, The Universityof Waikato]. https://researchcommons.waikato.ac.nz/handle/10289/15043.

(Hasan et al, 2018) Hasan, S. M. M., Mamun, M. A., Uddin, M. P., & Hossain, M. A. (2018). «Comparative Analysis of Classification Approaches for Heart Disease Prediction», in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, p. 14, doi: 10.1109/IC4ME2.2018.8465594.

(Helmbold et Schapire, 1995) Helmbold, D.P., Schapire, R.E. (1995). Predicting nearly as well as thebest pruning of a decision tree, eighth annual conference, ACM Press, pp. 61–68, https://doi.org/10.1145/225298.225305.

(Hewahi et Alashqar, 2015) Hewahi, N., & Alashqar, E. A. (2015). Wrapper Feature Selection based on GeneticAlgorithm for RecognizingObjects fromSatelliteImagery. Journal of Information Technology Research, 8 (3), 1-20.

(HongYang et al, 2020) C.-H. Yang, Y.-S.Chen, S.-H.Moi, L.-Y. Chuang, et Y.-D. Lin, « Identification of Kidney Clear Cell Carcinoma Mortality Risk-Associated Gene Mutation by Using a Random Survival Forest Approach », in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, oct. 2020, p. 61‑64.doi: 10.1109/BIBE50027.2020.00018.

(Houeland, 2011) Houeland, Tor Gunnar. 2011. « An Efficient Random Decision Tree Algorithm for Case-Based Reasoning Systems », 6.

(Hssina et al, 2014) Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5,Int.J.Adv.Comput.Sci, https://doi.org/10.14569/SpecialIssue.2014.040203.

(Huang et al, 2007) Huang, J., Cai, Y., & Xu, X. (2007). Ahybridgenetic algorithm for Feature selection wrapper based on mutual in formation.Pattern Recognition Letters, 28 (13), 1825-1844.

(Javeed et al, 2019) Javeed,A.,Zhou,S.,Yongjian,L., Qasim,I., Noor,A.,&Nour,R.(2019). «AnIntelligentLearningSystem Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection », IEEE Access, vol. 7, p. 180235‑180243, doi: 10.1109/ACCESS.2019.2952107.

(Kabiraj et al, 2020) S. Kabiraj et al., « Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm », in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, juill. 2020, p. 1‑4.doi: 10.1109/ICCCNT49239.2020.9225451.

(KASS, 1980) Kass, G.V., (1980).An Exploratory Technique for Investigating Large Quantities of Categorical Data.Appl.Stat.29, 119. https://doi.org/10.2307/2986296.

(Kaushik, 2016) Kaushik, S. (2016). Introduction to Feature Selection methods with an example (orhow to select the right variables? Disponible sur: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/.

(Khussainova et al, 2015) Khussainova, Gulmira, Sanja Petrovic, et Rupa Jagannathan. 2015. « Retrieval with Clustering in a Case-Based Reasoning System for Radiotherapy Treatment Planning ». Journal of Physics: Conference Series 616 (mai): 012013. https://doi.org/10.1088/1742-6596/616/1/012013.

(Kohavi et John, 1997) Kohavi, R. & John, G. H. (1997).Wrappers for feature subset selection. ArtificialIntelligence, 97 (1- 2), 273-324.

(Kolonder, 1993) Kolodner, J.L. (1993).Instructional Design: Case-Based Reasoning.13.

(Kuo et al, 2018) Kuo, C.Y., Yu, L.C., Chen, H.C., & Chan, C.L. (2018). «Comparison of Models for the Prediction of Medical Costs of Spinal Fusion in Taiwan Diagnosis-Related Groups by Machine Learning Algorithms», Healthc. Inform. Res., vol. 24, no 1, p. 29, doi: 10.4258/hir.2018.24.1.29.

(Leal et al, 2013) Leal, Yenny, Magda Ruiz, Carol Lorencio, Jorge Bondia, Luis Mujica, et Josep Vehi. 2013. «Principal Component Analysis in Combination with Case-Based Reasoning for Detecting Therapeutically Correct and Incorrect Measurements in Continuous Glucose Monitoring Systems ». Biomedical Signal Processing and Control 8 (6): 603‑14. https://doi.org/10.1016/j.bspc.2013.05.008.

.

(Lei et Yin, 2019) Lei, Zhang, et Dong Yin. 2019. « Intelligent Generation Technology of Sub-Health Diagnosis Case Based on Case Reasoning », 8.

(Lilhore et al, 2023) Lilhore, U.K., Manoharan, P., Sandhu, J.K. *et al.* Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. Sci Rep 13, 12473 (2023). https://doi.org/10.1038/s41598-023-36605-3.

(Li et al, 2020) Li, J., Tian, Y., Zhu, Y., Zhou, T., Li, J., Ding, K., & Li, J. (2020).Amulticenter Random forest model for effective prognosis prediction in collaborative clinical research network.Artificial Intelligence in Medicine, 103, 101814. https://doi.org/10.1016/j.artmed.2020.101814.

(Liu et al, 2017) Min, L. (2017). «An Improved Random Forest Method Base on RELIEFF for Medical Diagnosis», présentéà 2017 IEEE International Conference on Computational Scienceand Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA.

(Lounici et al, 2014) Nora, L., Hanya, K.M., Khadidja, S., (2014). Méthodes des arbres de décision pour le scoringbancaire, Economie et de Statistique Appliquée.

(Martinez et al, 2021) Marling, Cindy, ET Peter Whitehouse. 2001. « Case-Based Reasoning in the Care of Alzheimer's Disease Patients ». In Case-Based Reasoning Research and Development, édité par David W. Aha et Ian Watson, 2080:702‑15. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44593-5_50.

(Mishra et Suhas, 2016) Mishra, A., & Suhas, M. V. (2016). « Classificationof benignand malignant bonelesions onCT images using random forest », in 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, p. 1807‑1810,doi:10.1109/RTEICT.2016.7808146.

(Mohamed et al, 2012) Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H. (2012). A comparative study of Reduced Error Pruning method in decision tree algorithms, IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 392–397. https://doi.org/10.1109/ICCSCE.2012.6487177.

(Mohapatra et mohanty, 2018) Mohapatra, S. K., & Mohanty, M. N. (2018). «Analysis of Resampling Method for Arrhythmia Classification Using Random Forest Classifier with Selected Features», in 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, p. 495‑499,doi:10.1109/ICDSBA.2018.00098.

(Montani et al, 2000) Montani, Stefania, Riccardo Bellazzi, Luigi Portinale, Giuseppe d'Annunzio, Stefano Fiocchi, etMario Stefanelli. 2000. « Diabetic Patients Management Exploiting Case-Based ReasoningTechniques ». Computer Methods and Programs in Biomedicine 62 (3): 205‑18. https://doi.org/10.1016/S0169-2607(00)00068-7.

(Montani et al, 2003) S. Montani, P. Magni, R. Bellazzi, C. Larizza, A. V. Roudsari, et E. R. Carson, « Integrating model- based decision support in a multi-modal reasoning system for managing type 1 diabetic patients », Artif. Intell.Med. vol. 29, no 1‑2, p. 131‑151, sept. 2003, doi: 10.1016/S0933-3657(03)00045-9.

(Mustafa et al, 2023) Mustafa, E.M., Saad, M.M. & Rizkallah, L.W. Building an enhanced case-based reasoning and rule-based systems for medical diagnosis. J. Eng. Appl. Sci. 70, 139 (2023). https://doi.org/10.1186/s44147-023-00315-4.

(Nan et al, 2016) Nan, F., Wang, J., Saligrama, V. (2016). Optimally Pruning Decision Tree Ensembles with Feature Cost, arXiv: 1601.00955v1 [stat.ML], https://doi.org/10.48550/arXiv.1601.00955.

(Nasiri et Fathi, 2017) Nasiri, Sara, et Madjid Fathi. 2017. « Case Representation and Similarity Assessment in a Recommender System to Support Dementia Caregivers in Geriatric and Palliative Care», 10.

(Nguyen et al, 2013) Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). «Random forest classifié combined with feature selection for breast cancer diagnosis and prognostic», J. Biomed. Sci. Eng., vol.06,no05,p.551,560,doi:10.4236/jbise.2013.65070.

(Nouaouria, 2013) Nouaouria, N. (2013). Une approche d'optimisation par essaim de particules pour la recherche en mémoire de cas [Phdthesis]. UNIVERSITÉ DU QUÉBEC À MONTRÉAL**.**

(Perner, 1999) Perner, D. P. (1999). An Architecture for a CBR image segmentation system, Eng. Appl. Artif. Intell., vol. 12, p. 33, 1999.

(Perner et al, 2004) Perner, P., H. Perner, S. Janichen, et A. Buhring. 2004. « Recognition of Airborne Fungi Spores in Digital Microscopic Images ». In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 566-569 Vol.3. Cambridge, UK: IEEE. https://doi.org/10.1109/ICPR.2004.1334592.

(Pavithra et Geetha, 2022) M. Pavithra et B.T.Geetha, «Predictionof Chronic kidneycancer usingRBF supportvector machine compared with Random forest for better accuracy », in 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, juill. 2022, p. 15.doi: 10.1109/ICSES55317.2022.9914342.

(Proniewska et al, 2020) Proniewska, K., Pregowska, A., & Malinowski, K. P. (2020). «Identification of Human Vital Functions Directly Relevant to the Respiratory System Based on the Cardiac and Acoustic Parameters and Random Forest», IRBM, p. S1959031820300622, doi: 10.1016/j.irbm.2020.02.006.

(Quellec et al, 2008) Quellec, Gwenole, Mathieu Lamard, Guy Cazuguel, Christian Roux, et Beatrice Cochener. 2008. « Multimodal Medical Case Retrieval Using the Dezert-Smarandache Theory ». In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 394‑97.Vancouver, BC: IEEE. https://doi.org/10.1109/IEMBS.2008.4649173.

(Ramos et al, 2017) Ramos-González, Juan, DanielLópez-Sánchez, JoseA.Castellanos-Garzón, JuanF.DePaz, etJuan M. Corchado. 2017. « A CBR Framework with Gradient Boosting Based Feature Selection for Lung Cancer Subtype Classification ». Computers in Biology and Medicine 86 (juillet): 98‑106. https://doi.org/10.1016/j.compbiomed.2017.05.010.

(Raposo et al, 2020) Raposo, L. M., Rosa, P. T. C. R., & Nobre, F. F. (2020). «Random Forest Algorithm for Prediction of HIV Drug Resistance», in Pattern Recognition Techniques Applied to Biomedical Problems, M. R. Ortiz-Posadas, Éd. Cham: Springer International Publishing, p. 109‑127.doi:10.1007/978-3-030-38021-2_6.

(Rasovska, 2006) Ivana Rasovska. (2006), Contribution à une méthodologie de capitalisation des connaissances baséesur le raisonnement à partir de cas : Application au diagnostic dans une plateforme d'e-maintenance Thèse doctorale. L'UFR des Sciences et Techniques de l'Université de Franche-Comté.

(Razak et al, 2016) Razak,E., Yusof,F.,& Raus,R. A.(2016). «Classification of miRNA Expression Data Using Random Forests for Cancer Diagnosis», in 2016 International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, p. 187‑190, doi: 10.1109/ICCCE.2016.49.

(Saadi et Henni, 2019) Saadi, Fatima, Baghdad Atmani, et Fouad Henni. 2019. « Integration of Data mining Techniques into the CBR Cycle to Predict the Result of Immunotherapy Treatment ». In 2019 International Conference on Computer and Information Sciences (ICCIS), 1‑5. Sakaka, Saudi Arabia: IEEE.https://doi.org/10.1109/ICCISci.2019.8716415.

(Saeys et al, 2007) Saeys, Y., Inza, I., & Larranga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23 (19), 2507-2517.

(Sharaf-El-Deen et al, 2014) Sharaf-El-Deen, Dina A., Ibrahim F. Moawad, et M. E. Khalifa. 2014. « A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems ». Journal of Medical Systems 38(2):9.https://doi.org/10.1007/s10916-014-0009-1.

(Shi et al, 2020) W. Shi, B. Xiong, Y. Li, et M. Du, « Feature Selection of Input Variables for Diagnosis of Patellofe moral Pain Syndrome based on Random Forest and Multilayer Perceptron », in 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), Fuzhou, China, déc.2020,p.1,3.doi:10.1109/CSRSWTC50769.2020.9372597.

(Smiti et Elouedi, 2014) Smiti, Abir, et Zied Elouedi. 2014. « Case-Deletion Strategy for Maintaining the Case Based Reasoning System ». In 2014 Second World Conference on Complex Systems (WCCS), 37‑42. Agadir, Morocco: IEEE. https://doi.org/10.1109/ICoCS.2014.7060902.

(Speiser, 2021) Speiser, J.L. (2021). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J. Biomed. Inform. 117, 103763. https://doi.org/10.1016/j.jbi.2021.103763.

(Stracuzzi, 2007) Stracuzzi, D. J. (2007).Randomized Feature Selection. In: Liu, H., &Motoda, H. (eds) Computational Methods of Feature Selection. Chapman and Hall/CRC Press, 41-62.

(Sun et al, 2008) Sun, Zhaohao, Jun Han, et Dong Dong. 2008. « Five Perspectives on Case Based Reasoning ». In Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, édité par De-Shuang Huang, Donald C. Wunsch, Daniel S. Levine, et Kang-Hyun Jo, 5227:410‑19. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.https://doi.org/10.1007/978-3-540-85984-0_50.

(Tahmasebian et al, 2016) Tahmasebian, Shahram, Mostafa Langarizadeh, Marjan Ghazisaeidi, et and Mahdavi Mazdeh. 2016. « Designing and Implementation of Fuzzy Case-Based Reasoning System on Android Platform Using Electronic Discharge Summary of Patients with Chronic Kidney Diseases ». Acta Informatica Medica 24 (4): 266. https://doi.org/10.5455/aim.2016.24.266-270.

(Tarchoune et al, 2021) Tarchoune, I., Djebbar, A., Merouani, H.F., 2021. A Hybrid CBR Classification Model by integrating Decision Tree and Random Forest into Case Retrieval, International Conference on Networking and Advanced Systems (ICNAS), pp. 1–6, https://doi.org/10.1109/ICNAS53565.2021.9628920

(Tarchoune et al, 2022) Tarchoune, I., Djebbar, A., Merouani, H.F., Hadji, D., 2022. An Improved Random Forest Based on Feature Selection and Feature Weighting for Case Retrieval in CBR Systems: Application to Medical Data. Int. J. Softw. Innov, IJSI 10, 1–20, https://doi.org/10.4018/IJSI.293265.

(Tarchoune et al, 2023) Tarchoune, I., Djebbar, A., Merouani, H.F., 2023. A Case-Based Reasoning System-Based Random Forest for Classification: A Systematic Literature Review, Handbook of Research on Driving Socioeconomic Development With Big Data, IGI Global, pp. 170–196. https://doi.org/10.4018/978-1- 6684-5959-1.ch0.

(Tarchoune et al, 2024 a) Tarchoune, I., Djebbar, A., Merouani, H.F.D., Zenakhra, D.3FS-CBR-IRF: improvingcase retrieval for case-based reasoning with three feature selection and improved random forest.Multimed Tools Appl (2024).https://doi.org/10.1007/s11042-024-18360-3**.**

(Tarchoune et al, 2024) Tarchoune, I., Djebbar, A., Merouani, H.F, Harfi, R. A Novel Enhanced Random Forest for Medical Data Classification using Correlation Pearson and Best Number of Trees.Journal of Computing Science and Engineering (JCSE), vol. 18, no. 1, pp.57-68,March(2024).http://dx.doi.org/10.5626/JCSE.2024.18.1.57.

(Tharwat, 2018) Tharwat, A. (2018). Classification assessment methods.Applied Computing and Informatics, ahead- of-print (ahead-of-print). https://doi.org/10.1016/j.aci.2018.08.003.

(Tie et al, 2022) J. Tie, X. Lei, et Y. Pan, « Metabolite-disease association prediction algorithm combining DeepWalk and random forest », Tsinghua Sci. Technol., vol. 27, no 1, p. 58‑67, févr. 2022,doi:10.26599/TST.2021.9010003.

(Tyagi et Preetvanti, 2015) Tyagi, A et Singh, A. 2015. « ACS: Asthma Care Services with the Help of Case Base Reasoning Technique ». Procedia Computer Science 48: 561, 67. https://doi.org/10.1016/j.procs.2015.04.136.

(Vignesh et Revathy, 2019) Vignesh, S, D.R., & Revathy,R. (2019). «ADistincitve ModeltoClassify Tumor UsingRandomForest Classifier», inThird International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, p. 44,47, doi: 10.1109/ICISC44355.2019.9036473.

(Vijayakumari et Manikumaran, 2017) Vijayakumari, B., & Manikumaran, M. (2017). « PathologicalLungClassificationUsingRandomForestClassifier», p.5, 2017.

(VijiyaKumar et al, 2019) VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019). «Random Forest Algorithm forthe Prediction of Diabetes», inIEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, p. 1‑5,doi: 10.1109/ICSCAN.2019.8878802.

(Vorobieva et Schmidt, 2003) Vorobieva, Olga, Lothar Gierl, et Rainer Schmidt. 2003. « Adaptation Methods in an Endocrine Therapy Support System », 9.

(Wang et al, 2020) Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). «An improved random forest-based rule extraction method for breast cancer diagnosis», Appl. Soft Comput., vol. 86, p. 105941, doi: 10.1016/j.asoc.2019.105941.

(Wang et al, 2019) Wang, Y., Wang, D., Zhang, L., Liu, C., Li, J., Hou, F., & Peng, C.-K. (2019). Automaticidentification ofrapid eye movement sleep based onrandomforest usingheart ratevariability. Physica A: Statistical Mechanics and Its Applications, 527,121421. https://doi.org/10.1016/j.physa.2019.121421.

(Wang et al, 2022) Wang, Yameng, LiguoFei, YuqiangFeng, YanqingWang, et Luning Liu. 2022. «AHybridRetrieval Strategy for Case-Based Reasoning Using Soft Likelihood Functions ». Soft Computing 26 (7):3489‑3501.https://doi.org/10.1007/s00500-022-06733-5.

(Wardhani et al, 2022) Wardhani, A.K., Nugraha, E., Ulfiana, Q.(2022). Optimization of the Decision Tree Method using Pruning on Liver Disease Classification, J. Appl. Inform.Comput6, 136–140,https://doi.org/10.30871/jaic.v6i2.4350.

(Watson, 1997) Watson, I. (1997). Applying Case-Based Reasoning: Techniques for Enterprise Systems (1edition).MorganKaufmann.

(Weiss et Kulikowski, 1991) Weiss S.M. et Kulikowski C.A., Computer systems that learn. Artificial intelligence ISSN 0004-3702, Morgan-Kaufmann, 1991, 1993, Vol 62(2), pp.363-378, 1991.

(Williamson et al, 2022) Williamson, S., Vijayakumar, K., Kadam, V.J. (2022). Predicting breast cancer biopsy outcomes from BI-RADS findings using random forests with chi-square and MI features.Multimed. Tools Appl. 81, 36869–36889, https://doi.org/10.1007/s11042-021-11114-5.

(Wu et al, 2017) Wu, Y., Wang, H., & Wu, F. (2017). «Automatic classification of pulmonary tuberculosis andsarcoidosis based on random forest », in 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, p. 1‑5, doi:10.1109/CISP-BMEI.2017.8302280.

(Xu et al, 2017) Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X., & Zhu, T. (2017). «Cardiovascular risk prediction method based on CFS subset evaluationan d random forest classification framework», in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, p. 228‑232,doi:10.1109/ICBDA.2017.8078813.

(Yao et Li, 2022) B. Yao et C. Li, « Application Research on Identification of Cervical Lesions Based on Random Forest », in 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, janv. 2022, p. 1158‑1160.doi: 10.1109/ICPECA53709.2022.9719221.

(Yadav et al, 2024) Yadav, G., Bokhari, M.U. Hybrid Classifier for Optimizing Mental Health Prediction: Feature Engineering and Fusion Technique. Int J Ment Addiction (2024).Health https://doi.org/10.1007/s11469-024-01343-8.

(Yin et al, 2015) Yin, Ziming, Zhao Dong, Xudong Lu, Shengyuan Yu, Xiaoyan Chen, et Huilong Duan. 2015. « A Clinical Decision Support System for the Diagnosis of Probable Migraine and Probable Tension-Type Headache Based on Case-Based Reasoning ». The Journal of Headache and Pain16 (1):29.https://doi.org/10.1186/s10194-015-0512-x.

(Yu et al, 2020) Yu, Y., Wang, L., Huang, H., &Yang, W. (2020). «AnImprovedRandomForest Algorithm», J. Phys. Conf. Ser., vol. 1646, no 1, p. 012070, doi: 10.1088/1742-6596/1646/1/012070.

(Zemmal et al, 2019) Zemmal, N., Azizi, N., Ziani, A., Benzebouchi, N.E., Aldwairi, M. (2019). An Enhanced Feature Selection Approach based on Mutual Information for Breast Cancer Diagnosis, 6th International Conference on Image and Signal Processing and their Applications(ISPA), IEEE, pp.1–6.https://doi.org/10.1109/ISPA48434.2019.8966803.

(Zhang et al, 2021) Zhang, H., Shi, Y., Tong, J. (2021). Online supply chain financial risk assessment based on improved random forest. J. Data Inf. Manag. 3, 41–48. https://doi.org/10.1007/s42488-021-00042-6.

(Zhong et al, 2015) Zhong, S., Xie, X., & Lin, L. (2015). «Two-layer random forests model for case reuse in case-based reasoning», Expert Syst. Appl., vol. 42, no 24, p. 9412‑9425, doi: 10.1016/j.eswa.2015.08.005.

(Zhou et al, 2021) Zhou, H., Zhang, J., Zhou, Y., Guo, X., Ma, Y. (2021).A feature selection algorithm of decision tree based on feature weight. Expert Syst. Appl. 164, 113842. https://doi.org/10.1016/j.eswa.2020.113842.

(Zhu, 2007) Zhu, Z., Ong, Y.-S., Dash, M., (2007). Wrapper–Filter Feature Selection Algorithm Using a Memetic Framework. IEEE Trans. Syst. Man Cybern. Part B Cybern. 37, 70–76. https://doi.org/10.1109/TSMCB.2006.883267.

Productions scientifiques

Journals Paper

1. Tarchoune, I., Djebbar, A., Merouani, H.F., Hadji, D., 2022. An Improved Random Forest Based on Feature Selection and Feature Weighting for Case Retrieval in CBR Systems: Application to Medical Data. Int. J. Softw. Innov, IJSI 10, 1–20, https://doi.org/10.4018/IJSI.293265.

2. Tarchoune, I., Djebbar, A., Merouani, H.F.D., Zenakhra, D. 3FS-CBR-IRF: improving case retrieval for case-based reasoning with three feature selection and improved random Forest Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-18360-3.

3. Tarchoune, I., Djebbar, A., Merouani, H.F, Harfi, R. A Novel Enhanced Random Forest for Medical Data Classification using Correlation Pearson and Best Number of Trees.Journal of Computing Science and Engineering (JCSE), vol. 18, no. 1, pp.57-68, March (2024). http://dx.doi.org/10.5626/JCSE.2024.18.1.57.

Book Chapter

1. Tarchoune, I., Djebbar, A., Merouani, H.F., 2023. ACase-Based Reasoning System-Based Random Forest for Classification: A Systematic Literature Review, Handbook of Researchon Driving Socioeconomic Development With Big Data, IGI Global, pp. 170–196. https://doi.org/10.4018/978-1-6684-5959-1.ch008.

Conference Paper

1. Tarchoune, I., Djebbar, A., Merouani, H.F., 2021. A Hybrid CBR Classification Model by integrating Decision Tree and Random Forest into Case Retrieval, International Conference on Networking and Advanced Systems (ICNAS), pp. 1–6, https://doi.org/10.1109/ICNAS53565.2021.9628920.

2. Tarchoune, I., Djebbar, A., Merouani, H.F., 2023. Improving Random Forest with Pre-pruning technique for Binary classification, International Conference on Contemporary Academic Research (ICCAR), Konya, Turkey, Vol. 1 No. 2 (2023): AS-ABSTRACTS. https://doi.org/10.59287/as-abstracts.1202.

3. Tarchoune, I., Djebbar, A., Merouani, H.F., 2022. From Manuel to Automatic Feature Selection Applied in Random Forest, Journée scientifique des Mathématiques et de l'Informatique, (JSMI2022). Khemis Miiliana, Algeria. http://dx.doi.org/10.5626/JCSE.2024.18.1.57.