UNIVERSITÉ DE BLIDA 1
Faculté des Sciences
Département d'Informatique

THÈSE DE DOCTORAT

Option : Sciences Informatiques et de Données.

OPTIMISATION DE LA SELECTION DES CARACTÉRISIQUES VISUELLES POUR L'ANNOTATION D'IMAGES.

Par
Benkessirat Amina

devant le jury composé de :

N. Boustia	**Professeur, U. de Blida 1**	**Présidente**
N. Benblidia	**Professeur, U. de Blida 1**	**Directrice de thèse**
S.Oukid Khouas	**Maitre de conférence, U. de Blida 1**	**Examinatrice**
M. Farah	**Maitre de conférence, U. de Blida 1**	**Examinatrice**
Nait-Bahloul Safia	**Professeur, U. de Oran 1**	**Examinateur**

21 décembre 2022

Je dédie cette thèse à mes chers parents.

REMERCIEMENTS

Voilà, ma thèse est terminée. Une période assez particulière de ma vie vient de s'achever. Une période riche d'expériences et d'apprentissages scientifiques, mais aussi d'échanges humains instructifs. Si cette période s'est bien finie, c'est d'une part grâce au soutien des agréables personnes que j'ai pu côtoyer. Au regard de leurs qualités, mes remerciements semblent modestes. J'espère leur rendre ici une part de ce qu'ils méritent, tout en souhaitant que la vie me donne encore l'occasion d'affirmer ma gratitude envers eux.

En premier lieu, je remercie ALLAH, le tout puissant, de m'avoir donné la force pour survivre, ainsi que la volonté, la patience et le courage pour dépasser tous les obstacles et surmonter les moments de faiblesse.

Je suis grandement redevable à ma directrice de thèse Pr. Benblidia. Je tiens à la remercier de m'avoir offert l'opportunité de réaliser mon doctorat au sein du laboratoire LRDSI. Merci pour sa confiance, la liberté qu'elle m'a laissée pour explorer les pistes qui m'intéressaient, ses précieux conseils, son soutien inconditionnel et son appréciation positive. Sans ces paramètres, ce travail n'aurait pas été accompli. MERCI !

Je tiens à exprimer ma profonde gratitude à mon directeur de stage, au sein du laboratoire L2TI (Institut Galilée, Université Paris 13) Pr. Beghdadi pour son accueil, sa préoccupation de mon confort et mon avancement, ses précieux conseils et ses critiques constructives.

Je suis reconnaissante envers Allah, de m'avoir offert le morceau qui me complète et qui comble mes lacunes, ma soeur jumelle, mon demis Selma. Je ne la remercierai jamais assez pour tout simpelment ce qu'elle est et ce qu'elle fait de moi. Merci pour la valeur que tu ajoute à ma vie et à ma thèse, merci de me donner de ta lumière pour me voir briller. Merci pour ton courage et ta force malgrés les kilomètres que j'ai créés entre nous. MERCI !

Je tiens à remercier profondément Pr. Boustia, pour m'avoir fait l'honneur d'accepter de présider mon jury de thèse.

J'exprime toute ma gratitude aux membres du jury Pr. Bennouar, Dr. Farah et Dr. Boumahdi qui ont accepté de lire et évaluer mon modeste travail.

Je ne remercierai jamais assez mon ami Dr. Riali Isaak. Son aide et ses conseils dans chaque étape de mon parcours, du choix du sujets à la rédaction des articles et de la thèse ont été inestimables. Ses interventions m'ont été d'une grande aide dans la persévérance de mon travail. Isaak est mon modèle en tant que scientifique et conseiller.

Je tiens à remercier profondément mon amie Dr. Ykhlef Hadjer pour son soutien et son aide précieuse qui m'ont bien mis sur les railles de la recherche. Je la remercie infiniment pour son intervention lors des tests statistiques. Je lui en suis trés reconnaissante.

Je n'exprimerai jamais assez ma gratitude envers mes parents, chers Mama et Papa, pour tout leur soutien et amour. Merci pour leur confiance et la liberté qu'on m'a donnée ; ils ont planté une graine, ils l'ont arrosée et ils l'ont regardée pousser. Merci beaucoup pour tout. Je n'aurais pas pu m'épanouir et être ce que je suis sans eux et leurs prières.

Un grand merci à mes frères chéris Walid, Youcef et Imène pour leurs soutiens moral et de supporter la personne que je suis. Merci de me ramener l'eau à mon lit, de préparer ma tisane, de placer mon chargeur, de chercher mon chat quand il fugue et me faire des appels avec lui. Merci pour l'ambiance que vous me faite malgrès les kilomètres que j'ai créés entre nous. Toutes ces petites choses qui semblent banales font mon bonheur. Merci pour toute cette gâterie.

Des mots de gratitude particuliers vont à mon ami Allaoua qui a toujours été une source majeure de soutien lorsque les choses devenaient un peu décourageantes. Merci l'Ami d'être toujours là pour moi.

Je tiens à remercier le ministère de l'enseignement superieur et de la recherche scientifique pour la bourse octroyée.

Je tiens à remercier toute l'équipe du laboratoire L2TI, pour leur accueil chaleureux, leur confiance et leur ambiance familiale.

Je remercie également mes collègues et amis Mustafa et Fatima qui ont partagé avec moi des moments inoubliables à l'intérieur et à l'extérieur du laboratoire.

À tous ceux qui ont contribué de près ou de loin, MERCI.

Résumé

Un problème central de l'apprentissage automatique consiste à identifier un ensemble représentatif de caractéristiques à partir duquel sera construit un modèle de classification pour une tâche particulière. Le défi est d'identifier la relation entre ces caractéristiques de données et un certain point final, par exemple la classe cible pour une tâche de classification. Dans la plupart des ensembles de données, seules quelques caractéristiques sont pertinentes et contribuent à déterminer le point final. Les autres caractéristiques contribuent à la dimensionnalité globale de l'espace du problème, ce qui implique une mémoire importante pour stocker toutes les caractéristiques, un temps de traitement important pour obtenir le résultat souhaité, ou des résultats biaisés à cause des caractéristiques bruitées.

La sélection des caractéristiques est le processus de sélection automatique des caractéristiques importantes à partir des données. C'est une partie essentielle de l'apprentissage automatique, de l'intelligence artificielle, de l'exploration de données et de la modélisation. Il existe de nombreux algorithmes de sélection de caractéristiques disponibles et le choix approprié peut être difficile. L'objectif de cette thèse est de proposer une approche générique pouvant être appliquée dans différents domaines.

L'approche proposée est basée sur le calcul des valeurs propres avec des contraintes linéaires $Eigen\_FS$. Notre contribution a les caractéristiques suivantes : (a) elle prend à la fois en compte la pertinence des caractéristiques et la redondance par paire, (b) l'ensemble de départ n'est pas aléatoire et (c) elle est résolue par un algorithme itératif qui converge vers la solution optimale. Nous avons commencé par tester $Eigen\_FS$ sur des ensembles de données UCI ; les résultats expérimentaux révèlent l'efficacité de notre contribution par rapport à des méthodes de pointe. Par la suite, nous avons appliqué Eigen_FS au problème de la classification d'images endoscopiques ; dans le cadre de cette application, nous avons introduit un nouveau critère d'évaluation de caractéristiques. La détection des anomalies dans le tractus gastrointestinal par $Eigen\_FS$ a engendré une précision, sensibilité et spécificité de l'ordre de 98%, $95,3\%$ et $99,1\%$ (après la sélection des caractéristiques) contre $93,2\%$, 42% et $85,3\%$ respectivement (avant la sélection des caractéristiques). Les résultats atteints, attestent du fort potentiel du système décisionnel proposé, pour une aide au diagnostic clinique.

Table des matières

Table des figures

Liste des tableaux

List of Algorithms

Nomenclature

Acc	Accuracy (Exactitude)
BB	**B**ranch and **B**ound
CNN	**C**oncolutional **N**eural **N**etwork
DL	**D**eep **L**earning
DT	**D**ecision **T**ree
Err	**E**rror
F1	**F1**-score
FS	**F**eature **S**election
GI	**G**astro**i**ntestinal
KNN	**K** **N**earest **N**eighbors
Mc	**M**aximum **C**ut
MI	**M**utual **I**nformation
ML	**M**achine **L**earning
mRMR	**m**inimum **R**edundancy **M**aximum **R**elevance
Pr	**P**recision
Sn	**S**ensitivity
Sp	**S**pecificity
SVM	**S**upport **V**ector **M**achine
WCE	**W**ireless **C**apsule **E**ndoscopy
...	

Introduction Générale

Dans cette section, nous présentons le sujet de la sélection des caractéristiques, en expliquant l'importance du problème. Nous motivons ensuite le choix de la solution et énonçons les questions de recherche que nous aborderons dans les chapitres suivants. Ensuite, nous résumons les principales contributions réalisées au cours de la recherche doctorale. Enfin, nous concluons avec le plan de cette thèse et la liste des publications issues de notre travail recherche.

Motivation et orientation

En raison de la croissance explosive des technologies numériques, de plus en plus de données sont générées et stockées. L'avènement des données de haute dimension a posé des défis à la communauté des chercheurs en apprentissage automatique. En effet, la grande dimensionnalité rend la tâche d'apprentissage plus complexe et plus exigeante en termes de calcul et en termes d'espace. Le terme grande dimensionnalité s'applique à une base de données qui présente l'une des caractéristiques suivantes :

(a) le nombre d'échantillons est très élevé ;

(b) le nombre de caractéristiques est très élevé ;

(c) le nombre d'échantillons et de caractéristiques sont très élevés.

Le terme "grande dimensionnalité" fait l'objet d'une certaine controverse dans la littérature. Certains auteurs affirment qu'il ne se réfère qu'à l'espace des caractéristiques, tandis que d'autres l'utilisent indistinctement pour les deux types de caractéristiques susmentionnées. Dans cette thèse, la troisième alternative sera adoptée. Lorsque les algorithmes d'apprentissage traitent des données de haute dimension, ils peuvent voir leurs performances amoindries en raison d'un surajustement. Les modèles formés perdent de leur interprétabilité car ils sont plus complexes, par conséquent la vitesse et l'efficacité des algorithmes diminuent en fonction de la taille de l'espace des caractéristiques.

Le défi est d'identifier la relation entre ces caractéristiques de données et un certain point final, par exemple la classe cible pour une tâche de classification. Dans la plupart des ensembles de données, seules quelques caractéristiques sont pertinentes et contribuent à déterminer le point final. Les autres caractéristiques contribuent à la dimensionnalité globale de l'espace du problème, ce qui implique une mémoire importante pour stocker toutes les caractéristiques, un temps de

traitement important pour obtenir le résultat souhaité, ou des résultats biaisés à cause des caractéristiques bruitées. Par conséquent, le prétraitement des données est d'une grande importance. La sélection d'un sous ensemble de caractéristiques est considérée comme une solution appropriée pour résoudre les problèmes ci-dessus et est devenue une composante primordiale du processus d'apprentissage automatique, trouvant son succès dans de nombreuses applications différentes du monde réel, en particulier celles liées aux problèmes de classification.

La sélection des caractéristiques constitue une partie essentielle de l'apprentissage automatique, de l'intelligence artificielle, de l'exploration de données et de la modélisation. Le processus de sélection élimine ou identifie automatiquement les caractéristiques redondantes, non pertinentes ou hautement corrélées de l'ensemble de données de grande dimension sans changer la sémantique du problème. Autrement dit, la sélection des caractéristiques est le processus qui vise à trouver un sous-ensemble "pertinent" de caractéristiques à partir de l'ensemble de départ, qui décrit correctement le problème donné avec une dégradation minimale ou même une amélioration des performances.

En effet, la sélection des caractéristiques pertinentes avant la tâche d'apprentissage améliore souvent la précision de prédiction des modèles d'apprentissage, réduit les temps d'apprentissage et de prédiction des modèles d'apprentissage et aboutit à des modèles d'apprentissage simples. Dans cette thèse, nous nous sommes intéressées à l'étude du processus de la sélection des caractéristiques, dans un cadre général. Tout d'abord, une analyse critique des méthodes de sélection de caractéristiques existantes est effectuée, afin de vérifier leur adéquation aux différents défis et de pouvoir fournir des recommandations aux utilisateurs. A l'issue de cette analyse, nous avons proposé une nouvelle approche de sélection de caractéristiques. Notre approche est générique et peut être appliquée pour résoudre des problèmes de sélection de caractéristiques provenant de différents domaines d'application.

Objectifs de la recherche

Les approches de sélection de caractéristiques évaluent les caractéristiques individuellement en attribuant un poids à chaque caractéristique, ou évaluent un sous-ensemble de caractéristiques en utilisant une stratégie de recherche [1, 3]. Pour sélectionner les caractéristiques pertinentes, l'espace de recherche contient évidemment 2n-1 candidats, pour un vecteur de n caractéristiques. Il a été prouvé que ce problème est NP-difficile [1, 4], et il semble complexe de produire une solution globalement optimale. L'obtention d'une précision de prédiction élevée avec de faibles coûts de calcul reste un défi.

Il existe trois catégories générales de techniques de sélection de caractéristiques, celles basées sur [5] :

— *le clustering :* les caractéristiques avec des propriétés informatives similaires sont regroupées ; chaque cluster est minimisé séparément pour réduire la taille globale de l'ensemble.

— *le classement :* les caractéristiques de l'ensemble sont ordonnées en fonction de certaines mesures d'évaluation prédéfinies (en prenant en compte la pertinence et la redondance), les caractéristiques du sous-ensemble sont ensuite sélectionnées en fonction de cet ordre.

— *l'optimisation :* la sélection des caractéristiques peut être considérée comme un problème d'optimisation combinatoire dans le but de trouver un sous-ensemble qui optimise un critère prédéfini qui est une estimation de la performance prédictive.

Contrairement aux trois catégories mentionnées ci-dessus, en utilisant une formulation mathématique, on peut modéliser la performance du sous-ensemble d'une manière plus fondée sur des principes. Malheureusement, comme discuté ci-dessus, résoudre le problème d'optimisation globalement est difficile. Par conséquent, des algorithmes génétiques [6] et des relaxations semi-définies [7] ont été proposés pour résoudre approximativement le problème. Cependant, bien que la complexité temporelle de ces méthodes ne soit plus exponentielle, il y a toujours le problème d'une faible scalabilité [8, 9].

Dans le cadre de cette thèse, nous proposons une nouvelle approche basée sur l'optimisation pour résoudre le problème de sélection de caractéristiques. L'objectif est d'optimiser efficacement un critère qui intègre à la fois la pertinence et la redondance. Contrairement aux algorithmes génétiques et aux relaxations semi-définies discutés ci-dessus, nous formulons le problème comme un problème d'optimisation de calcul des valeurs propres avec des contraintes linéaires, qui peut être résolu avec un algorithme efficace dont la convergence globale est garantie. En effet, l'approche proposée est hybride : elle assure la fonction d'optimisation et de classement. Elle pallie ainsi au problème majeur des approches d'optimisation et de classement.

Notre approche a les caractéristiques suivantes : (a) elle prend en compte, à la fois, la pertinence des caractéristiques et la redondance par paires, (b) l'ensemble de départ n'est pas aléatoire et (c) elle est résolue par un algorithme itératif qui converge vers la solution optimale.

Contributions

— Notre première contribution [1] est un état de l'art des méthodes de sélection de caractéristiques. Nous avons établi une classification et discussion des différents travaux existants afin d'en tirer les limites. L'étude de l'existant nous a permis de réaliser notre contribution principale.

— Notre deuxième et principale contribution [5] consiste à améliorer la sélection en se basant sur l'optimisation combinatoire en général, et le calcul des valeurs propres $Eigen\_FS$. Notre contribution vise à améliorer l'exactitude de la classification en général. Elle tient compte de la pertinence des caractéristiques et de la redondance. Afin de tester la robustesse de notre approche, nous avons appliqué $Eigen\_FS$ sur 20 ensembles de donnees UCI.

— Notre troisième contribution [10] est une initiation à la classification des images endoscopiques en utilisant un schéma classique de l'apprentissage automatique. Il s'agit d'une

classification en utilisant l'apprentissage par transfert dans un réseau de neurones profond et pré-entrainé. Le travail qui a été fait nous a permis de comprendre le problème de la dimensionnalité des images médicales, en particulier celui des images endoscopiques.

— Notre quatrième contribution est une application directe de notre deuxième contribution *Eigen_FS* sur des images endoscopiques, pour faire une classification classique. Cette contribution comprend tous les modules d'un système d'aide au diagnostic médical en commençant par le prétraitement jusqu'à atteindre la classification finale. Notre approche générique a permis la classification d'images endoscopiques pour la détection des anomalies dans le tracus gastrointestinal

Organisation de la thèse

La thèse est composée de six chapitres qui sont organisés comme suit :

— **Chapitre 1 « Sélection de caractéristiques » :** dans ce chapitre, nous définissons en détail le processus de la sélection des caractéristiques. Ensuite, nous présentons les méthodes de sélection de base.

— **Chapitre 2 « Apprentissage et classification automatique : un aperçu»** : dans le deuxième chapitre, nous nous concentrons sur les systèmes de classification les plus utilisés par la communauté des chercheurs. Nous commençons par la classification par apprentissage machine, ensuite la classification en utilisant le deep learning. Nous clôturons ce chapitre par une proposition d'un modèle profond pour la classification des images médicales.

— **Chapitre 3 « Sélection de caractéristiques pour la classification** » : dans le troisième chapitre, nous présentons quelques travaux de la littérature. Nous classons les contributions par domaine d'application.

— **Chapitre 4 : « Sélection des caractéristiques basée sur le calcul des valeurs propres avec contraintes » :** dans ce chapitre, nous présentons une nouvelle approche de sélection de caractéristiques, basé sur le calcul des valeurs propres avec contraintes. Nous commençons ce chapitre par des notions mathématiques de base pour la conception de notre approche. Nous finissons par détailler notre contribution.

— **Chapitre 5 : « Résultats expérimentaux» :** dans ce chapitre, nous détaillons le protocole expérimental suivi pour la validation de notre approche. Nous présentons également une comparaison équitable de notre approche proposée avec une approche de sélection de base, à savoir *mRMRM* et deux approches récentes.

— **Chapitre 6 : « Classification des images endoscopiques en utilisant** $Eigen\_FS$ **pour la sélection des caractéristique visuelle » :** dans ce chapitre, nous appliquons notre méthode proposée *Eigen_FS* sur des données médicales. Nous présentons en détail tout la chaine de classification permettant d'interpréter automatiquement les images endoscopiques.

La thèse est clôturée par une conclusion générale qui inclut quelques pistes de recherche pour l'amélioration de notre approche.

Chapitre 1

Sélection de caractéristiques

1.1 Introduction

Le défi de la tâche d'exploration de données est d'identifier la relation entre les caractéristiques des données et un certain point final, par exemple la classe cible pour une tâche de classification [1]. C'est un outil de prétraitement important dans l'exploration de données et est devenu un domaine actif de recherche et de développement depuis quelques décennies [11]. Dans la plupart des ensembles de données, seules quelques caractéristiques sont pertinentes et contribuent à déterminer le point final. Les autres caractéristiques contribuent à la dimensionnalité globale de l'espace du problème, ce qui implique une mémoire importante pour stocker toutes les caractéristiques, un temps de traitement important pour obtenir le résultat souhaité, ou des résultats biaisés à cause des caractéristiques bruitées [1, 12]. Par conséquent, le prétraitement des données est d'une grande importance. La sélection d'un sous-ensemble de caractéristiques est considérée comme une solution appropriée pour résoudre les problèmes ci-dessus et est devenue une composante primordiale du processus d'apprentissage automatique [1]. Ce processus élimine ou identifie automatiquement les caractéristiques redondantes, non pertinentes ou hautement corrélées entre elles de l'ensemble de données de haute dimension sans changer la sémantique du problème. Autrement, la sélection de caractéristiques est le processus qui vise à trouver un sous-ensemble "pertinent" de caractéristiques à partir de l'ensemble de départ. La pertinence des caractéristiques dépend des objectifs du système [1, 11, 12]. Trois qualificatifs peuvent classer une caractéristique selon sa pertinence [1] : i) une caractéristique non pertinente doit être éliminée de l'ensemble original. ii) une caractéristique f est faiblement pertinente s'il existe un sous-ensemble V où la performance de V est meilleure que celle de $V \cup f$. iii) une caractéristique f est fortement pertinente si son absence du sous-ensemble sélectionné implique une détérioration significative de la performance du système. Les approches de sélection de caractéristiques évaluent les caractéristiques individuellement en attribuant un poids à chaque caractéristique, ou évaluent un sous-ensemble de caractéristiques en utilisant une stratégie de recherche [1, 3]. Pour sélectionner les caractéristiques pertinentes, l'espace de recherche contient évidemment $2n-1$

candidats. Il a été prouvé que ce problème est NP-hard [1, 4], et il semble complexe de produire une solution globalement optimale. Les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection sont classées en trois grandes catégories principales [1, 12] : ”filter”, ”wrapper” et ”embedded”. A travers ce chapitre, nous allons :

— Définir les notions, concepts et procédures de base de la sélection de caractéristiques.

— Décrire les techniques de sélection de caractéristiques de référence.

— Identifier les problèmes existants de la sélection de caractéristiques et proposer des moyens de les résoudre.

Dans ce qui suit, nous allons introduire de façon détaillée le processus de la sélection des caractéristiques et nous allons présenter quelques méthodes de pointe. Dans un chapitre ultérieur, nous allons passer en revue quelques approches de sélection de caractéristiques selon le domaine d'application.

1.2 Pertinence des caractéristiques

Cette section est consacrée aux définitions de la *pertinence* qui ont été proposées dans la littérature [1, 12, 13, 14]. Le sous-ensemble final doit contenir toutes les caractéristiques pertinentes. Par conséquent, la pertinence des caractéristiques doit être correctement définie en fonction de leur contribution à la solution. Dans la littérature, trois qualificatifs décrivent les caractéristiques : non pertinente, faiblement pertinente et fortement pertinente. Soit F l'ensemble complet des caractéristiques, f_i une caractéristique, $S_i = F - \{f_i\}$ et C l'étiquette de classe en supposant un problème de classification. Ces trois qualificatifs sont formalisés comme suit :

— **Non pertinente :** une caractéristique f_i est non pertinente si et seulement si :

$$\forall S_i' \subseteq S_i, P(C|f_i, S_i') = P(C|S_i)$$

— **Faiblement pertinente :** une caractéristique f_i est faiblement pertinente, si et seulement si :

$$P(C|f_i, S_i) = P(C|S_i)$$

et

$$\exists S_i', telque P(C|f_i, S_i') \neq P(C|S_i)$$

— **Fortement pertinente :** une caractéristique f_i est fortement pertinente, si et seulement si :

$$P(C|f_i, S_i) \neq P(C|S_i)$$

De nombreuses définitions ont été proposées dans la littérature pour répondre à la question "pertinente à quoi ?" [15].

— **Pertinente à un concept cible :** une caractéristique f_i est pertinente pour un concept cible c s'il existe une paire d'instances A et B dans l'espace des instances telle que A et B ne diffèrent que par leur affectation à f_i et $c(A) \neq c(B)$.
Une autre façon d'énoncer cette définition est que la caractéristique f_i est pertinente s'il existe une instance dans l'espace d'instances pour laquelle la modification de la valeur de f_i affecte la classification donnée par le concept cible.

— **Fortement pertinent pour l'échantillon/la distribution :** Une caractéristique f_i est fortement pertinente pour l'échantillon S s'il existe des instances A et B dans S qui ne diffèrent que par leur affectation à f_i et ont des étiquettes différentes. De même, f_i est fortement pertinente pour la cible c et la distribution D s'il existe des instances A et B ayant une probabilité non nulle sur D qui ne diffèrent que par leur affectation à f_i et satisfont $c(A) \neq c(B)$.

— **Peu pertinente pour l'échantillon/la distribution :** la caractéristique f_i est faiblement pertinente pour l'échantillon S (ou pour la cible c et la distribution D) s'il est possible de supprimer un sous-ensemble de caractéristiques de sorte que f_i devienne fortement pertinente.

— **Utilité incrémentale :** étant donné un échantillon de données S, un algorithme d'apprentissage L et un ensemble de caractéristiques E, la caractéristique f_i est progressivement utile à L par rapport à E si la précision de l'hypothèse que L produit en utilisant l'ensemble de caractéristiques $f_i \cup E$ est meilleure que la précision obtenue en utilisant uniquement le jeu de fonctionnalités E.

1.3 Le processus de la sélection des caractéristiques

De façon générale, étant donné $F = \{(f_1, f_2, ..., f_N)\}$ un ensemble de N caractéristiques et Ev une fonction d'évaluation d'un sous-ensemble de caractéristiques. Nous supposons que le meilleur ensemble de caractéristiques engendre la plus grande valeur de Ev. L'objectif de la sélection de caractéristiques est de trouver un sous-ensemble $F' \subseteq F$, de taille $N' \leq N$, tel que :

$$Ev(F') = \max_{Z \subseteq F} Ev(Z) \tag{1.1}$$

Tel que $|Z| = N'$ et N' est un nombre défini par l'utilisateur ou contrôlé par une méthode de génération de sous ensemble, que nous décrivons dans les prochaines sous-sections. D'après Dash et Liu [16] la procédure de la sélection des caractéristiques passe par quatre étapes clés : génération du sous-ensemble, évaluation du sous-ensemble, critères d'arrêt et validation des résultats, comme résumé dans la **Figure 1.1**.

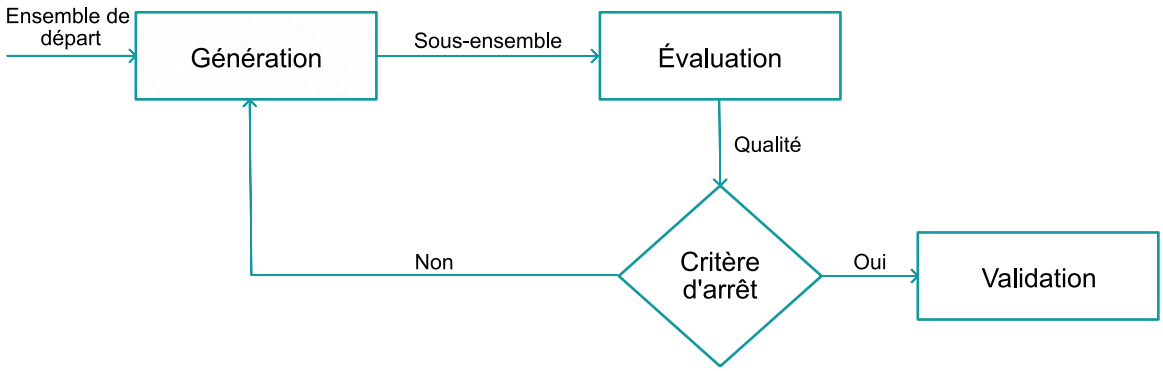

FIGURE 1.1 – Procédure générale d'un algorithme de sélection de caractéristiques

1.3.1 Génération du sous ensemble

Avant d'appliquer un algorithme de recherche, il faut définir un point de départ [1, 12, 16]. Dans le cadre de notre étude, le point de départ est l'ensemble des caractéristiques d'entrée. L'ensemble de départ peut être vide et différentes caractéristiques peuvent être ajoutées successivement ou inclure toutes les caractéristiques et certaines d'entre elles sont supprimées si nécessaire. En outre, l'ensemble de départ peut être un sous-ensemble de caractéristiques sélectionnées au hasard et une ou plusieurs caractéristiques sont ajoutées ou supprimées à chaque itération. Une fois l'ensemble de départ bien défini, une stratégie de recherche est choisie. Le choix de l'algorithme de recherche est important pour sélectionner les caractéristiques pertinentes de façon efficace. Pour un espace de N caractéristiques, il existe $2^N - 1$ sous-ensembles candidats [1, 12, 16]. Cet espace de recherche est massif, avec une valeur modérée de N une recherche approfondie et exhaustive est impossible. Par conséquent, trois stratégies sont proposées dans la littérature : exponentielle, séquentielle et aléatoire.

1.3.1.1 Stratégie exponentielle

La recherche exponentielle n'est pas nécessairement exhaustive, mais c'est une recherche optimisée qui garantit la meilleure solution. Différentes fonctions heuristiques peuvent être utilisées pour réduire l'espace de recherche sans échapper la solution optimale. Par exemple, BRANCH AND BOUND [17] et Beam Search [18] sont évalués pour un plus petit nombre de sous-ensembles et garantissent un sous-ensemble optimal. La complexité de ces algorithmes est de l'ordre de 2^N.

1.3.1.2 Stratégie séquentielle

Dans les procédures séquentielles, une ou plusieurs caractéristiques peuvent être ajoutées ou supprimées séquentiellement, plusieurs variantes de cette stratégie ont été proposées dans la littérature, à savoir [1, 12, 16] :

— **Forward** ou ascendante, la recherche commence à partir d'un ensemble vide et au moins une caractéristique est ajoutée à chaque itération.

— **Backward** ou descendante, contrairement à la variante "Forward", la recherche commence à partir d'un ensemble contenant toutes les caractéristiques et au moins une d'entre elles est supprimée à chaque itération.

— **Stepwise** ou par étape, cette variante mélange les variantes ci-dessus et ajoute ou supprime au moins une caractéristique à chaque itération.

Les variantes *forward, backward et stepwise* sont facile à mettre en œuvre car leur coût est polynomial et est de l'ordre N^2 ou moins [1]. Cependant, cette stratégie ne garantit pas un résultat optimal, car la solution optimale peut être dans une région de l'espace de recherche qui n'est pas visitée.

1.3.1.3 Stratégie aléatoire

Les algorithmes de cette stratégie génèrent un nombre fini de sous-ensembles et le meilleur est ensuite sélectionné. Tous les algorithmes de cette catégorie utilisent l'aléatoire pour échapper à l'optimum local dans l'espace de recherche qui est de l'ordre de N^2 [1].

1.3.2 Evaluation du sous ensemble

L'évaluation du pouvoir discriminant d'une caractéristique est une étape importante du processus de sélection. Dans le domaine de l'exploration de données, il existe trois approches générales pour effectuer la sélection : "filter", "wrapper" et "embedded". Ces méthodes sont utilisées pour l'évaluation, en se basant sur différents critères. Les critères d'évaluation sont divisés en deux catégories : les critères dépendants et les critères indépendants [1, 12]. Les critères dépendants sont utilisés par des procédures d'évaluation enveloppantes "wrapper", pour évaluer les sous-ensembles en se basant sur un algorithme d'apprentissage. Généralement, ils génèrent de meilleurs résultats que les critères indépendants, mais ils sont plus coûteux en termes de temps de calcul et de ressources utilisées [1, 12, 19].Les critères indépendants sont utilisés par des procédures d'évaluation filtrantes "filter" pour évaluer les sous-ensembles sans impliquer un algorithme d'apprentissage.

1.3.2.1 Approches d'évaluation

— *Filter :* cette catégorie de méthode est considérée comme un prétraitement de données, puisque les caractéristiques sont évaluées indépendamment de l'algorithme d'apprentissage. Les algorithmes de filtrage sont divisés en deux catégories, univariées et multivariées [1, 12]. Les caractéristiques des algorithmes univariés sont évaluées une par une, sans tenir compte de leur dépendance avec les autres caractéristiques [1, 12, 19]. Ces méthodes permettent d'éliminer les caractéristiques qui semblent inutiles mais qui pourraient être utiles si elles sont combinées avec d'autres, ce qui affecte négativement les performances du système. Afin de résoudre ce problème, des méthodes multivariées sont proposées

[1, 12, 19]. A la fin de l'évaluation, un score de pertinence est attribué à chaque caractéristique. Ce score mesure le pouvoir discriminatif de chaque caractéristique par rapport au problème donné [20]. Un seuil est défini pour supprimer les caractéristiques inutiles [20]. L'ensemble restant constitue l'entrée de l'algorithme résolvant le problème [16, 20]. Les approches filtrantes sont efficaces et rapides à calculer [1]. La démarche générale d'une approche "filter" est résumée dans **l'Algorithme 1**. Les avantages des techniques de filtrage sont [21] :

— Adaptation facile à des ensembles de données de très haute dimension.

— Simple et rapide à calculer.

— Indépendance de l'algorithme d'apprentissage.

Algorithm 1: Algorithme des approches "filter"

Input :
D={X, L} // Ensemble de données d'apprentissage, tel que $X = \{f_1, f2, ..., f_n\}$ et L les labels.
X' // Ensemble initial de caractéristiques, tel que X'=∅ ou X' $\subseteq$ X.
θ // Critère d'arrêt. **Output :**
X'_{opt} ; //Ensemble optimal
Begin

Initialisation
X_{opt}=X' ;
$\rho_{opt} = E(X', I_m)$; // Evaluer X' en utilisant un critère indépendant I_m
X_g=générer(X) ; // Générer un sous ensemble pour évaluation
$\rho = E(X_g, I_m)$; // Evaluation de X_g par I_m
If ($\rho \geq \rho_{opt}$)
$\rho_{opt} = \rho$; $X'_{opt} = X_g$
repeat (tant que θ n'est pas atteint) ;
End.
Retourner X'_{opt}
End.

La procédure du modèle "filter" est illustrée par la**figure 1.2**.

— *Wrapper* : dans cette catégorie, la recherche d'un sous-ensemble pertinent est basée sur un algorithme d'apprentissage comme fonction d'évaluation. Ils se divisent en deux catégories : déterministe ou aléatoire. Il semble que cette approche ait de meilleures performances par rapport aux approches filtrantes [12, 22]. Cependant, ces méthodes présentent deux inconvénients principaux [22]. Le sous-ensemble sélectionné dépend fortement du

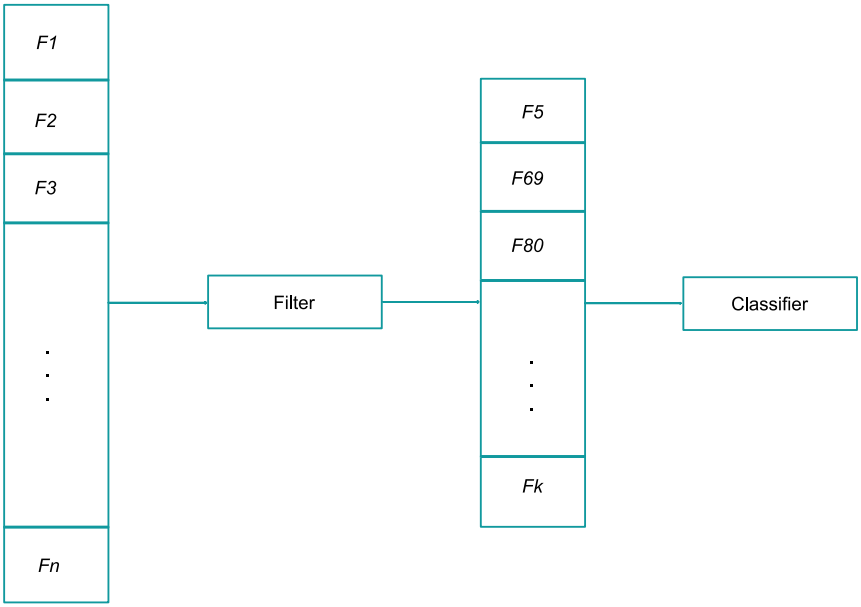

FIGURE 1.2 – La procédure du modèle ”filter”

model d'apprentissage et la complexité de calcul est élevée. Cette complexité rend impossible l'utilisation d'une stratégie exhaustive (problème NP-complet) [12, 20]. La démarche générale d'une approche "wrapper" est résumée dans **l'Algorithme 2**.

Algorithm 2: Algorithme des approches "wrapper"

Input :
D={X, L} // Ensemble de données d'apprentissage, tel que $X = \{f_1, f2, ..., f_n\}$ et L les labels.
X' // Ensemble initial de caractéristiques, tel que X'=∅ ou X' ⊆ X.
θ // Critère d'arrêt.
Output :
X'_{opt} ; //Ensemble optimal
Begin
//Initialisation
X_{opt}=X' ;
$\rho_{opt} = E(X', A)$; // Evaluer X' en utilisant un algorithme d'apprentissage A
X_g=générer(X) ; // Générer un sous ensemble pour évaluation
$\rho = E(X_g, A)$; // Evaluation de X_g par A
Si ($\rho \geq \rho_{opt}$) **alors**
    $\rho_{opt} = \rho$
    $X'_{opt} = X_g$
repeat (tant que θ n'est pas atteint)
End
Retourner X'_{opt}
End.

La procédure du modèle "wrapper" est illustrée par la **figure 1.3**.

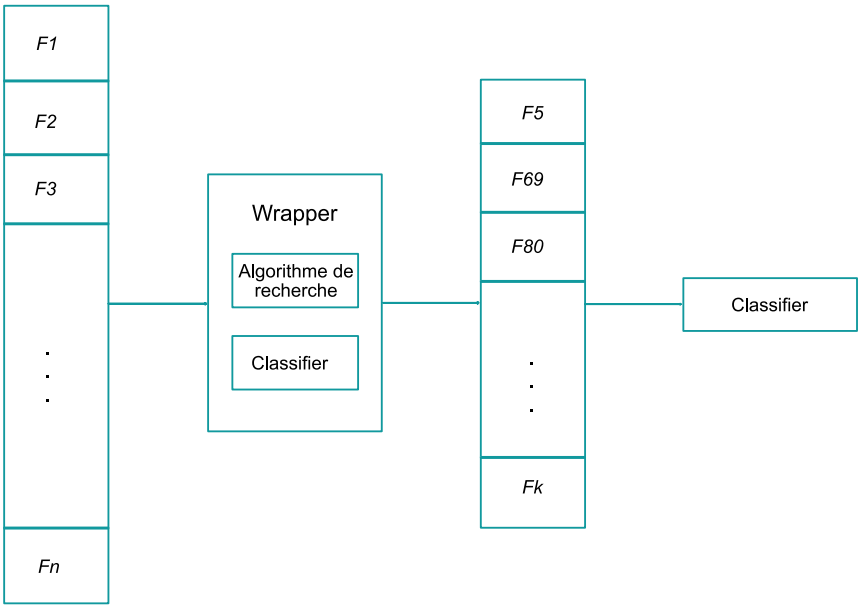

FIGURE 1.3 – La procédure du modèle "wrapper"

— *Embedded :* cette approche interagit avec l'algorithme d'apprentissage nécessitant un coût de calcul inférieur à celui de l'approche wrapper [1, 12]. Elle capture également la dépendance des caractéristiques. Elle prend en compte non seulement la corrélation entre une caractéristique d'entrée et la caractéristique de sortie, mais recherche également localement les caractéristiques qui permettent une meilleure discrimination locale. Elle utilise des critères indépendants pour décider des sous-ensembles optimaux pour une cardinalité connue. Ensuite, l'algorithme d'apprentissage est utilisé pour sélectionner le sous-ensemble optimal final parmi les sous-ensembles optimaux pour différentes cardinalités. La démarche générale d'une approche "embedded" est résumée dans **l'Algorithme 3**.

Algorithm 3: Algorithme des approches "embedded"

Input :
D={X, L} // Ensemble de données d'apprentissage, tel que $X = \{f_1, f2, ..., f_n\}$ et L les labels.
X' // Ensemble initial de caractéristiques, tel que X'=$\emptyset$ ou X' $\subseteq$ X.
θ // Critère d'arrêt.
Output :
X'_{opt} ; //Ensemble optimal
Begin

Initialisation
X_{opt}=X' ;
$\rho_{opt} = E(X', I_m)$; // Evaluer X' en utilisant un critère indépendant I_m
$\sigma_{opt} = E(X', A)$; // Evaluer X' en utilisant un algorithme d'apprentissage A
C_0=C(X') ; //Calculer la cardinalité de X'
Pour (k=C_0+1 ; k $\leq$ n ; k++) faire
    X_g=$X_{opt} \cup \{f_i\}$; // Générer sous ensemble de cardinalité k, pour évaluation
    $\rho = E(X_g, I_m)$; // Evaluer X_g en utilisant un critère indépendant I_m
    Pour (i=0 ; i $\leq$ n-k ; i++) faire
        Si $\rho \geq \rho_{opt}$ **alors**
            $\rho_{opt} = \rho$;
            $X'_{opt} = X_g$
        End
        $\sigma_{opt} = E(X'_{opt}, A)$ // Evaluer X'_{opt}avec un algorithme d'apprentissage A
    Si $\sigma \geq \sigma_{opt}$ **alors**
        $X'_{opt} = X'_{opt}$
        $\sigma_{opt} = \sigma$
        Sinon
            Break and return
    End
End
Retourner X'_{opt}
End.

1.3.2.2 Critères d'évaluation

Les sous ensemble générés doivent être évalués par des critères d'évaluation, qui sont classés dans deux catégories, critères indépendants et critères dépendants. Dans ce qui suit, nous allons introduire quelques critères dépendants et indépendants.

— *Critères indépendants :* fondamentalement, les modèle "filter" utilisent des critères indépendants pour évaluer les sous-ensembles des caractéristiques générés, sans impliquer aucun modèle d'apprentissage. De nombreux critères indépendants sont proposés dans la littérature, à savoir les mesures de distance [23], les mesures d'information ou d'incertitude [24], les mesures de probabilité d'erreur [25], les mesures de dépendance [26], les mesures de dépendance interclasse [12] et les mesures de consistance [27].

— **Mesures de distance :** ce critère est connu sous le nom de divergence ou discrimination et séparabilité, qui calcule la divergence ou la distance probabiliste entre les densités de probabilité conditionnelles des classes. Dans le problème à deux classes, la meilleure caractéristique est celle qui fait la plus grande différence entre les probabilités conditionnelles de classes. Soit S l'espace de toutes les distributions de probabilité et $D : S \times S \longrightarrow \mathbb{R}$ une divergence sur S. Alors, étant données p et q les distributions de probabilité continues, la condition suivante doit être satisfaite :

$$\begin{cases} D(p,q) \geq 0, \forall p,q \in S \\ D(p,q) = 0 sinon \\ La matrice g(D) est strictement d finie positive sur S \end{cases}$$

Le sous-ensemble de caractéristiques $X' \subset X$ est choisi comme solution, si leur divergence entre les probabilités conditionnelles est significative. En d'autres termes, les probabilités d'une caractéristique faiblement pertinente sont très similaires. Quelques mesures sont présentées ci-dessous [12, 28, 29] :

$$Kullback\_Libler : D_{KL}(p,q) = \int p(x) \log_2 (\frac{p(x)}{q(x)} dx$$

$$Bhattacharya : D_B(p,q) = \int \sqrt{p(x) \times q(x)} dx$$

$$Jeffrey's divergence : D_j(p,q) = \int (p(x) - q(x)) \log_2 (\frac{p(x)}{q(x)}) dx$$

$$Matusita : D_M(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$$

$$Kagan's\ divergence : D_K(p,q) = \frac{1}{2} \int \frac{((p(x) - q(x))^2}{p(x)} dx$$

— **Mesure d'information ou incertitude :** cette mesure est basée sur le gain d'information des caractéristiques. Le gain d'information d'une caractéristique est défini comme la différence entre l'incertitude antérieure et l'incertitude postérieure attendue. Le gain d'information est maximal pour des classes de probabilité égale, et

l'incertitude est minimale. L'entropie de *Shannon* est largement utilisée pour les mesures d'incertitude et est définie comme suit :

$$H(A) = -\sum_{i=1}^{n}(p_i \log_2 p_i)$$

Tel que A est un espace de probabilité discret.

— **Mesure de probabilité d'erreur :** cette mesure consiste à minimiser la probabilité d'erreur. Considérons deux catégories de cas tels que les classes C_1 et C_2 qui se divisent en deux régions R_1 et R_2, de manière éventuellement non optimale. Les erreurs de classification peuvent se produire dans les deux régions : R_1, l'état réel de la nature est C_1 et $R2$, l'état réel de la nature est C_2. Ces événements sont mutuellement exclusifs. La probabilité de l'erreur est calculée comme suit :

$$P(erreur) = \int_{R_2}(p(x/C_1)p(C_1)dx + \int_{R_1}(p(x/C_2)p(C_2)dx$$

En générale, si $(p(x/C_1)p(C_1) > (p(x/C_2)p(C_2)$, x doit être classifier dans R_1. C'est le concept qui permet de réaliser la règle de décision bayésienne.

— **Mesure de dépendance :** c'est une mesure classique. Cette mesure est également connue sous le nom de mesure de similarité ou de corrélation. Il existe deux grandes catégories de mesures qui peuvent être utilisées pour mesurer la corrélation entre deux variables aléatoires. L'une est basée sur la corrélation linéaire classique et l'autre sur la théorie de l'information. Parmi ces deux catégories, la mesure la plus connue est le coefficient de corrélation linéaire. Si deux caractéristiques sont linéairement dépendantes, leur coefficient de corrélation est de 1, si non le coefficient de corrélation est égal à 0. Dans le contexte de la sélection des caractéristiques pour la classification, les caractéristiques fortement corrélées sont préférées. Selon la littérature standard, pour une paire de variables (X,Y), le coefficient de corrélation linéaire $'r'$ est calculé comme suit :

$$r = \frac{\sum((X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum(X_i - \overline{X_i})^2}\sqrt{\sum(Y_i - \overline{Y_i})^2}}$$

Dans la littérature, plusieurs mesures de corrélation ont été proposées à partir de la formule classique [30].

— **Mesure de consistance :** cette mesure tente de trouver un nombre minimum de caractéristiques qui séparent les classes de manière aussi cohérente que l'ensemble complet de caractéristiques. Tel qu'une incohérence est définie comme deux instances ayant les mêmes valeurs de caractéristiques mais des étiquettes de classe différentes [12, 31].

— **Mesures de dépendance interclasse**

— *Critères dépendants :* ils sont utilisés par les modèles "wrapper" et exige un algorithme d'apprentissage pour la sélection des caractéristiques ; tel qu'il utilise la performance de l'algorithme d'apprentissage appliqué sur le sous-ensemble sélectionné pour déterminer quelles caractéristiques sélectionner. Les critères dépendants donnent généralement des performances supérieures car ils trouvent les caractéristiques les mieux adaptées à l'algorithme d'apprentissage qui résout un le problème donné, mais ils ont également tendance à être plus coûteux en termes de temps de calcul et peuvent ne pas convenir à d'autres algorithmes d'apprentissage [12, 31]. Par exemple, dans le cadre de l'apprentissage supervisé, plus particulièrement dans une tâche de classification, la précision prédictive est largement utilisée comme mesure primaire. Elle peut être utilisée comme un critère dépendant pour la sélection des caractéristiques. Comme les caractéristiques sont sélectionnées par le classifieur qui utilise ensuite ces caractéristiques pour prédire les étiquettes de classe des instances non visitées, la précision est normalement élevée, mais l'estimation de la précision pour chaque sous-ensemble de caractéristiques est plutôt coûteuse en termes de calcul [13]. Un autre exemple, dans le cadre de l'apprentissage non supervisé, plus particulièrement dans une tâche de clustering, le modèle wrapper tente d'évaluer la qualité d'un sous-ensemble de caractéristiques par la qualité des clusters résultant de l'application de l'algorithme de clustering sur le sous-ensemble sélectionné. Il existe un certain nombre de critères heuristiques pour estimer la qualité des résultats du clustering, tels que la compacité du cluster, la séparabilité de la dispersion et le maximum de vraisemblance. Des travaux récents sur le développement de critères dépendants dans la sélection de caractéristiques pour le clustering peuvent être trouvés dans [32, 33, 34]

1.3.3 Critère d'arrêt

Les techniques de sélection nécessitent un critère d'arrêt pour éviter une recherche exhaustive des sous-ensembles. Plusieurs critères d'arrêt sont proposés dans la littérature, principalement [1, 12, 35] :

— La recherche est terminée.

— Un nombre maximal d'itérations prédéfini est atteint.

— Le nombre maximal prédéfini de caractéristiques est atteint.

— L'ajout ou la suppression d'une caractéristique ne produit pas un meilleur sous-ensemble.

— Un sous-ensemble optimal est trouvé.

1.3.4 Validation des résultats

Il existe deux alternatives de validation qui dépendent de la nature des données utilisées [1, 12, 36]. Si nous connaissons à l'avance les caractéristiques pertinentes comme dans le cas des données synthétiques, la validation est faite directement en comparant cet ensemble connu

de caractéristiques avec les caractéristiques sélectionnées. Cependant, dans les applications du monde réel, nous n'avons généralement pas une telle connaissance préalable. Dans ce cas, la précision du sous-ensemble sélectionné est testée en utilisant un classifieur. Si nous utilisons le taux d'erreur de classification comme indicateur de performance pour une tâche d'exploration, pour un sous-ensemble de caractéristiques sélectionné, nous pouvons simplement mener l'expérience «avant-après» pour comparer le taux d'erreur du classifieur appris sur l'ensemble complet de caractéristiques et celle apprise sur le sous-ensemble sélectionné [31].

1.4 Domaines d'application dans le monde réel

Dans tous les domaines d'application, la collecte des données est une étape clé. Cependant, le nombre considérable de caractéristiques constituant ces données, spécialement les caractéristiques non pertinentes et redondantes restent un problème majeur dans la collecte de données. C'est pourquoi plusieurs méthodes de sélection de caractéristiques ont été développées et appliquées avec succès dans différents domaines [1, 12], tels que [1, 37] :

— Recherche et indexation d'image.

— Analyse génomique.

— Détection d'intrusion.

— Téledetection.

— Classification du text.

Le chapitre 3 sera consacré à l'étude de quelques méthodes de ces différents domaines.

1.5 Méthodes de sélection des caractéristiques les plus utilisées

Cette section est consacrée à la présentation des méthodes de sélection des caractéristiques les plus utilisées, basées sur différentes techniques.

1.5.1 Méthodes de sélection séquentielles

— *SFS (Sequential Forward Selection)* : cette méthode a été proposée en 1963 par Marill et Green [38]. La sélection séquentielle avant est une recherche ascendante qui commence par un ensemble vide et ajoute une nouvelle caractéristique à chaque itération. Formellement, elle ajoute la caractéristique candidate f_i qui maximise $I(S;C)$ au sous-ensemble de caractéristiques sélectionnées S, c'est-à-dire :

$$S = S \cup \{\arg\max_{f_i \in F \setminus S}(I(S, f_i); C)\}$$

La première version simple de la sélection séquentielle croissante ne permet pas de prendre en considération la dépendance entre les variables. Même si, une caractéristique peut être

non pertinente individuellement, mais pertinente si elle est combinée avec d'autres caractéristiques. Ceci est prouvé par la sélection séquentielle arrière.

— *SBS (Sequential Backward Selection)* : cette méthode a été proposée en 1971 par Whitney [39]. La sélection séquentielle arrière est une recherche descendante qui commence par l'ensemble de toutes les caractéristiques et supprime une à chaque itération. Formellement, il s'agit de supprimer les caractéristiques les moins informatives, à chaque itération :

$$S = S \backslash \{\arg\min_{f_i \in S}(I(S \backslash f_i); C)\}$$

Habituellement, SBS est plus coûteuse en termes de calcul que SFS. Cependant, elle est plus efficace, puisqu'elle prend en considération l'interaction entre les caractéristiques.

— *BDS (Bidirectional Search)* : c'est une fusion des deux méthodes, SFS et SBS [22]. À chaque itération des caractéristiques sont ajoutées et des caractéristiques sont retirées dans les deux directions jusqu'à ce que l'ensemble des caractéristiques sélectionnées soit atteint quelle que soit la direction. Ceci permet de réduire le temps de calcul par rapport à la sélection séquentielle arrière pour des ensembles de grande taille.

1.5.1.1 minimum-Redundancy, Maximum-Relevance (mRMR)

C'est une méthode de référence, proposée en 2005 par Peng et autre [40]. Leur algorithme prend en considération à la fois la maximisation de la pertinence et ma minimisation de la redondance entre les caractéristiques. Etant donné un ensemble de départ, de N instances et M caractéristiques (pour chaque instance), $X = \{x_i; i = 1...M\}$ et le label de class C, mRMR est définie comme suit :

$$max\Phi(D,R);\ \Phi = D - R$$

Tel que, le critère de maximisation de la pertinence est calculé comme suit :

$$D = maxD(S,c)$$

$$D = \frac{1}{\mid S \mid} \sum_{x_i} I(x_i, c)$$

Et le critère de minimisation de la redondance est calculé comme suit

$$R = minR(S)$$

$$R = \frac{1}{\mid S \mid^2} \sum_{x_i, x_j} I(x_i, x_j)$$

Cet algorithme définit la redondance d'une variable x_i dans S par la valeur moyenne des informations mutuellement échangées, entre elle et x_j, tel que $i, j \in S$ et $i \neq j$. D'un autre côté, cet

algorithme calcule la pertinence d'une variable $x_i \in S$ par la moyenne des valeurs des informations mutuelles entre la variable elle et l'ensemble des étiquettes de classes c.

1.5.2 ReliefF

Relief est un algorithme proposé en 1992 par Kira et Rendell [41], il est considéré comme un algorithme de prétraitement de données. Le critère de pertinence est utilisé pour évaluer et classer les caractéristiques, sans prendre en compte la redondance. Ensuite en 2003 par Marko et Kononenko, il a été amélioré et adapté au cas des multi-classes, sous le nom de ReliefF [42]. La procédure générale est constituée de 5 étapes :

1. Sélectionner une instance quelconque I_m de la classe C.
2. Rechercher dans le voisinage H (des instances appartenant à C), les k plus proches.
3. Rechercher les k plus proches instances, de l'espace M (n'appartenant pas à C).
4. Mettre à jours l'estimation w_i de la caractéristique f_i, basé sur I_m, H et M.
5. Mettre à jour w_i, en utilisant la formule suivante :

$$w_i = w_i - \frac{\sum_{k=1}^{k} D_H}{n_c \times k} + \sum_{c=1}^{C-1} P_c \times \frac{\sum_{k=1}^{k} D_{Mc}}{n_c \times k}$$

Tel que, n_c c'est le nombre d'instance appartenant à la classe c, D_H c'est la somme de la distance entre les instances sélectionnées et les instances de l'espace H, D_M c'est la somme de la distance entre les instances sélectionnées et l'espace M et P_c c'est la probabilité à priori de la classe c.

1.5.3 Branch and Bound

Branch and Bound (BB) est un algorithme proposé en 1977 par Narendra et Fukunaga [43] qui considère une représentation graphique du problème de la recherche des meilleures caractéristiques. Le graphe représentant le problème est construit comme suit :

— La racine représente l'ensemble des caractéristiques.

— Les nœuds fils représentent des sous-ensembles de caractéristiques.

L'algorithme parcourt l'arbre de la racine jusqu'aux feuilles, et au fur et à mesure élimine la moins pertinente des caractéristiques du noeud courant, celle-ci ne satisfait pas le critère de sélection. Un seuil est défini (bound) pour supprimer les sous arbres d'un nœud, si la valeur de ce nœud est inferieur au seuil. L'algorithme BB garantit de trouver un sous-ensemble optimal de caractéristiques à si la fonction d'évaluation utilisée est monotone. Cependant, le temps de calcul croissant avec l'augmentation du nombre des caractéristiques est un grand inconvénient. En 2003 Chan et Xue [44] propose une amélioration en utilisant d'autres techniques de recherche dans l'arbre afin d'accélérer le processus de sélection.

1.6 Conclusion

Dans ce chapitre, nous avons défini en détail les différents modules de la sélection des caractéristiques, à savoir : la génération du sous ensemble, l'évaluation du sous ensemble, les critères d'arrêt et la validation. La génération du sous ensemble est essentielle pour commencer l'évaluation et peut être exponentielle, séquentielle ou aléatoire. L'évaluation du sous ensemble peut être une approche "filter", "wrapper" ou "embedded". Les critères d'arrêt peuvent être des critères par rapport à la taille du sous ensemble final ou par rapport à la pertinence du sous ensemble final. La validation des résultats se fait dépendamment de la nature des données (réelles ou synthétiques). Nous avons également étudié la pertinence des caractéristiques. Dans le chapitre suivant, nous allons présenter les critères d'évaluation d'un algorithme de sélection des caractéristiques dans le cadre de la classification. Également, nous passons en revue quelques méthode de sélection qui ont été développées et testées en utilisant des données réelles ou synthétiques.

Chapitre 2

Apprentissage et classification automatique : un aperçu

2.1 Introduction

La classification est une procédure permettant de classer des instances (images, textes, signaux...) en plusieurs catégories, en fonction de leurs similitudes. Nous pouvons facilement comprendre ou analyser notre environnement en classifiant les objets en fonction de notre objectif. Mais il n'est pas toujours facile de les classer, surtout lorsqu'ils contiennent du bruit ou informations non pertinente. Dans le cadre de ce travail de recherche, nous nous intéressons particulièrement à la classification d'images.

La classification d'images est toujours une tâche critique mais importante pour de nombreuses applications, y compris les systèmes d'aide au diagnostic médical. Il est parfois très difficile d'identifier un objet dans une image. En particulier lorsqu'elle contient du bruit, des parasites ou une mauvaise qualité. Si une image contient plus qu'un objet, cette tâche devient plus difficile. Donc on peut dire que le principe principal de la classification d'image est de reconnaître les caractéristiques de cette dernière.

Il existe trois principales techniques de classification, à savoir la classification supervisée, la classification non supervisée et la classification semi-supervisée [45]. Chaque classification est précédée par d'autres étapes, notamment l'extraction et la sélection des caractéristiques.

Dans ce chapitre, nous allons définir les différentes techniques de classification. Ensuite, nous présentons les meilleurs modèles de classification, d'après "IEEE International Conference on Data Mining (ICDM)". Par la suite, nous présentons la classification profonde et nous finissons par la présentation d'un nouveau modèle de classification profonde, en utilisant l'apprentissage par transfert.

2.2 Techniques de classification

Comme nous avons dit précédemment, il existe trois techniques de classification, appartenant à trois techniques d'apprentissage, à savoir supervisé, non supervisé et semi-supervisé. Dans ce qui suit nous allons présenter brièvement ces trois techniques.

2.2.1 Classification supervisée

Dans l'apprentissage supervisé, on utilise des points de données étiquetés. Donc la phase d'apprentissage est nécessaire dans la classification supervisé [45]. La classification supervisée est donc menée par l'apprentissage supervisé. Soit une variable d'entrée X, une variable de sortie Y et $Y = f(X)$ une fonction de mappage de l'entrée à la sortie [46]. Le but de l'apprentissage est d'approximer la fonction de mappage si bien que lorsque de nouvelles données d'entrée (X) sont présentées, les variables de sortie (Y) sont prédites pour ces données. C'est ce qu'on appelle l'apprentissage supervisé parce que le processus d'apprentissage d'un algorithme à partir de l'ensemble de données d'apprentissage peut être considéré comme un supervise le processus d'apprentissage. Nous connaissons les bonnes sorties ; l'algorithme fait itérativement des prédictions sur les données d'apprentissage et est corrigé par les données préalablement connues.

2.2.2 Classification non supervisée

Deuxièmement, la classification non supervisée n'utilise pas de données étiquetées, ce qui signifie que la phase d'apprentissage n'est pas supervisée [45]. L'apprentissage non supervisé est une technique d'apprentissage automatique. Le modèle est conçu pour fonctionner seul et découvrir des informations. Les algorithmes sont laissés à eux-mêmes pour découvrir et présenter la structure intéressante des données. L'apprentissage non supervisé est très utile dans l'analyse exploratoire car il peut automatiquement identifier la structure des données [46]. Par exemple, si un analyste essayait de segmenter les consommateurs, les méthodes de regroupement non supervisées seraient un excellent point de départ pour son analyse. Dans les situations où il est impossible ou peu pratique pour un humain de proposer des tendances dans les données, l'apprentissage non supervisé peut fournir des informations initiales qui peuvent ensuite être utilisées pour tester des hypothèses individuelles. La classification par apprentissage non supervisé est appellée "clustering".

2.2.3 Classification semi-supervisée

La plus grande différence entre l'apprentissage automatique supervisé et non supervisé est la suivante : les algorithmes d'apprentissage automatique supervisé sont formés sur des ensembles de données étiquettées qui guident l'algorithme pour comprendre quelles fonctionnalités sont importantes pour le problème à résoudre [45, 46]. Il s'agit d'un processus très coûteux, en particulier lorsqu'il s'agit de gros volumes de données. Les algorithmes d'apprentissage automatique

non supervisés, en revanche, sont formés sur des données non étiquetées et doivent déterminer eux-mêmes l'importance des fonctionnalités en fonction des modèles inhérents aux données. L'inconvénient de tout apprentissage non supervisé est que son spectre d'application est limité. Pour contrer ces inconvénients, le concept d'apprentissage semi-supervisé a été introduit [46]. Dans ce type d'apprentissage, l'algorithme est formé sur une combinaison de données étiquetées et non étiquetées. Typiquement, cette combinaison contiendra une très petite quantité de données étiquetées et une très grande quantité de données non étiquetées. Ceci est utile pour plusieurs raisons. Premièrement, le processus d'étiquetage de quantités massives de données pour l'apprentissage supervisé prend souvent beaucoup de temps et coûte cher. De plus, trop d'étiquetage peut imposer des biais humains au modèle. Cela signifie que l'inclusion de nombreuses données non étiquetées pendant le processus de formation tend en fait à améliorer la précision du modèle final tout en réduisant le temps et les coûts consacrés à sa construction.
Pour résumer, les techniques d'apprentissage non supervisé sont utilisées pour découvrir et apprendre la structure des variables d'entrée. Les techniques d'apprentissage supervisé sont utilisées pour faire des prédictions optimales pour les données non étiquetées, réintroduire ces données dans l'algorithme d'apprentissage supervisé en tant que données d'apprentissage et utiliser le modèle pour faire des prédictions sur de nouvelles données invisibles.

2.3 Les meilleurs algorithmes de classification

Dans cette section nous présentons quelques modèles de classification qui appartiennent aux meilleurs 10 algorithmes d'exploration de données, d'aprés IEEE *International Conference on Data Mining (ICDM)* en Décembre 2006, à savoir : C4.5, k Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, et CART. Ces dix meilleurs algorithmes figurent parmi les algorithmes d'exploration de données les plus influents dans la communauté des chercheurs [5, 47].

2.3.1 Arbres de décision

Decision trees DT sont efficaces pour la prise de décision [5]. Ils sont exprimés comme une partition récursive de l'espace d'instance [48]. Un DT est constitué de nœuds qui forment un arbre à racines, c'est-à-dire un arbre dirigé avec un nœud appelé "racine" qui n'a pas d'arêtes entrantes [47, 48]. Tous les autres nœuds ont exactement un bord entrant. Un nœud avec des bords sortants est appelé nœud interne ou nœud de test. Tous les autres nœuds sont appelés feuilles qui sont les nœuds de décision. Dans un DT, chaque nœud interne divise l'espace d'instance en deux ou plusieurs sous-espaces en fonction d'une certaine fonction discrète des valeurs des attributs d'entrée. Dans le cas le plus simple et le plus fréquent, chaque test prend en compte un seul attribut, de sorte que l'espace d'instance est partitionné en fonction de la valeur de l'attribut. Dans le cas d'attributs numériques, la condition fait référence à un intervalle. Chaque feuille est affectée à une classe représentant la valeur cible la plus appropriée. Alternativement,

la feuille peut contenir un vecteur de probabilité indiquant la probabilité que l'attribut cible ait une certaine valeur. Les instances sont classées en les faisant naviguer de la racine de l'arbre vers une feuille, en fonction du résultat des tests effectués le long du chemin.

2.3.2 Machines à vecteurs de support

Support Vector Machine SVM sont des modèles d'apprentissage automatique incontournables et sont considérés comme des modèles très robustes [5]. Les chercheurs montrent que les SVMs utilisent un espace de haute dimension pour trouver un hyperplan afin d'effectuer une classification binaire où le taux d'erreur est minimal [49]. Le problème des SVMs est de séparer les deux classes avec une fonction obtenue à partir des données d'apprentissage disponibles [47, 49]. L'objectif est de produire des classificateurs qui généralisent d'autres problèmes. Les vecteurs d'entrée sont maximaux pour séparer deux régions qui sont la fonction hyperplan des SVMs. Cependant, SVMs sont insensibles au nombre de dimensions, donc ils ne sont pas limités à la séparation de deux types d'objets. Il y a plusieurs alternatives aux lignes de division qui arrangent l'ensemble des objets en deux classes. Cette technique cherche à trouver une fonction classificatrice optimale qui peut séparer deux ensembles de données de deux catégories différentes. Dans ce cas, la fonction de séparation visée est linéaire :

$$g(x) = sing(f(x))$$

Tel que $f(x) = w^T x + b; w, x \in \mathbb{R}^n et b \in \mathbb{R}$. w et b sont les paramètres pour lesquels on recherche une valeur. Le meilleur hyperplan est situé au milieu de deux ensembles d'objets de deux classes. Trouver le meilleur hyperplan équivaut à maximiser la marge ou la distance entre deux ensembles d'objets de deux classes. Les échantillons situés le long d'un hyperplan sont appelés vecteurs de support.

2.3.3 K-plus proches voisins

K-nearest neighbor knn est l'un des plus simple classifieurs et des plus triviaux [16]. Ce modèle se repose sur l'hypothèse que les instances de chaque classe sont entourées principalement d'instances de la même classe [50]. De ce fait, l'algorithme commence par un ensemble d'instances d'apprentissage dans l'espace des caractéristiques et un scalaire k. Une nouvelle instance est classée en lui attribuant l'étiquette la plus fréquente parmi les k échantillons d'apprentissage les plus proches de cette instance. Plusieurs distances peuvent être utilisées pour mesurer la distance entre les instances, la plus utilisée est la distance euclidienne [50, 51]. La distance euclidienne est calculée par la formule suivante [51] :

$$L(x_i, x_j) = (\sum_{i,j=1}^{n} (|x_i - x_j|^2))^{1/2}$$

2.4 Métriques d'évaluation

La performance des modèles de classification est évaluée sur la base du calcul de plusieurs métriques, la plus importante est l'exactitude, plus souvent dite l'accuracy [1, 5]. L'accuracy est le nombre de fois où le modèle formé est correct. Un système de classification doit classer correctement toutes les instances de l'ensemble de données, mais les performances d'un système de classification ne sont pas entièrement exemptes d'erreurs. L'erreur d'un système de classification consiste à classer de nouveaux objets dans une classe incorrecte, c'est ce qu'on appelle donc erreur de classification (Error rate). Pour évaluer un système de classification, on se base essentiellement sur le résultat de la matrice de confusion. La matrice de confusion est un tableau enregistrant les résultats de la classification et est représentée dans le tableau **2.1**.

TABLE 2.1 – Matrice de confusion [1, 2]

Class original	Class predite	
	Classe positive	Classe negative
Classe positive	TP	FN
Classe negative	FP	TN

Tel que :

— TP (true positif) c'est le vrai positif.

— TN (true negatif) c'est le vrai négatif.

— FP (false positif) c'est le faux positif.

— FN (false negatif) c'est le faux négatif.

Le TP est une condition lorsque les observations provenant de classes positives sont prédites comme étant positives. Le TN est une condition lorsque les observations provenant de classes négatives sont prédites comme étant négatives. Le FP est une condition lorsque l'observation réelle provient de classes négatives mais est prédite comme étant positive. Le FN est une condition lorsque l'observation réelle provient d'une classe positive mais prédite négative.

L'évaluation des performances d'un système de classification peut être justifiée par l'accuracy (Acc), la précision (Pr) et le rappel (Rec). Le taux de rappel/TP peut être défini comme le niveau de précision des prédictions dans les classes positives, puis le pourcentage du nombre de prédictions qui sont justes sur les observations positives. De plus, l'accuracy est le pourcentage des prédictions globales qui sont justes sur toutes les observations du groupe de données.

En se basant sur le contenu de la matrice de confusion, nous calculons Acc, Pr, Rec et Err comme suit :

$$Acc = \frac{TN + TP}{TP + FP + FN + TN}$$

$$Pr = \frac{TP}{TP+FP}$$

$$Rec = \frac{TP}{TP+FN}$$

$$Err = \frac{FP+FN}{TP+FP+TN+TF}$$

D'autres mesures peuvent être calculées à travers la matrice de conclusion, elles sont définies ci-dessous. F_Measure (FM) représente la moyenne harmonique entre les valeurs de rappel et de précision et se calcule comme suit :

$$FM = \frac{2 \times Pr \times Rec}{Pr + Rec}$$

La mesure de sensibilité (sensitivity Sn) est utilisée pour mesurer la fraction d'instances positives qui sont correctement classés. La mesure de spécificité (Specificity Sp) est utilisée pour mesurer la fraction d'instances négatives qui sont correctement classés. Elles sont calculées comme suit :

$$Sn = \frac{TP}{TP+FN}$$

$$Sp = \frac{TN}{TN+FP}$$

La moyenne géométrique (Geometric-Mean GM) est utilisée pour maximiser le taux de TP et le taux de TN, tout en maintenant simultanément les deux taux relativement équilibrés. Elle se calcule comme suit :

$$GM = \sqrt{TP \times TN}$$

Outre que l'examen de la matrice de confusion, l'évaluation de la qualité de la prédiction d'un classificateur peut se faire à partir des courbes ROC (Receiver Operating Characteristic) [2],AUC (Area Under the Curve) et la mesure de Kappa de Cohen (Ka) [2].
La mesure d'évaluation Kappa de Cohen (Ka) est une mesure permettant de déterminer la fiabilité ou le niveau de similitude de deux variables ou plus. Elle est calculée comme suit :

$$Ka = \frac{p_0 - p_e}{1 - p_e}$$

Tel que p_0 c'est la proportion totale de la diagonale principale de la fréquence d'observation, p_e c'est proportion totale marginale de la fréquence d'observation. La valeur Ka peut être interprétée avec la force de l'accord :

- Mauvais $\leq$ 0,20.
- Moyen = [0,21-0,40].
- Modéré =[0,41-0,60].
- Bon = [0,61-0,80].
- Très bon = [0,81-1,00].

2.5 Comparaison des modèles de classification

Dans cette section nous allons présenter quelques tests statistiques qui sont utilisés pour l'évaluation et la comparaison des classifieurs, suivant la recommandation de Demšar [5, 52].

2.5.1 Le test de Friedman

C'est un test non paramétrique qui est considéré comme équivalent du test paramétrique ANOVA à mesures répétées. Les algorithmes sont classés séparément pour chaque ensemble de données, l'algorithme le plus performant obtenant le rang 1, le second le rang 2. . . [5]. En cas de corrélation entre les données, des rangs moyens sont attribués [52]. Soit r_{ij} le rang du j^{me} des k algorithmes sur le i^{me} des N ensembles de données. Le test de Friedman compare les rangs moyens des algorithmes :

$$R_j = \frac{1}{N}\sum r_i^j$$

En supposant que tous les systèmes ont des performances similaires, donc que leurs rangs R_j devraient être égaux, la statistique :

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

suit la distribution du Chi-squared avec un degré de liberté de (k - 1) pour N et k suffisamment grands (généralement N > 10 et t > 5). Pour un plus petit nombre d'algorithmes et d'ensembles de données, des valeurs critiques exactes ont été calculées [53]. Iman et Davenport [54] ont montré que le χ_F^2 de Friedman est conservatrice et ont dérivé une meilleure statistique :

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)\chi_F^2}$$

qui est distribué selon la distribution F avec degrés de liberté de $k-1$ et $(k-1)(N-1)$. Le tableau des valeurs critiques peut être trouvé dans n'importe quel livre de statistique. Ce test permet uniquement d'évaluer si les différences observées dans les performances sont statistiquement significatives. Afin d'avoir une vue plus détaillée de ce à quoi ces différences correspondent précisément, c'est-à-dire identifier les paires de techniques dont les performances sont significativement différentes, nous effectuons généralement un test post hoc lorsque le test de Friedman rejette l'hypothèse nulle. Nemenyi, Bonfferoni-Dunn, et Holm sont des exemples de tests post hoc qui sont largement utilisés en conjonction avec le test de Friedman.

2.5.2 Le test de Nemenyi

Ce test est invoqué lorsque toutes les techniques sont comparées entre elles. Les performances de deux méthodes sont significativement différentes si leurs rangs moyens correspondants diffèrent d'au moins la différence critique :

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

où la valeur critique q_α est définie sur la base de la statistique de la gamme studentisée divisée par $\sqrt{2}$.

2.5.3 Le test de Bonferroni-Dunn

En général, le test de Bonferroni-Dunn est conservateur et a peu de puissance ; néanmoins, ce test est utile lorsque nous sommes uniquement intéressés par la comparaison de toutes les techniques avec un algorithme de contrôle. Dans ce cas spécifique, le test de Bonferroni-Dunn est plus puissant que le test de Namenyi car ce dernier ajuste la valeur critique pour faire des comparaisons $k(k-1)$, alors que lorsque l'on compare avec une méthode de contrôle, on ne fait que des comparaisons $k-1$. Ce test est fondamentalement défini de manière similaire au test de Nemenyi, sauf que nous estimons la valeur critique pour le niveau de signification $\alpha/(k-1)$.

2.6 Classification en utilisant l'apprentissage profond

Dans cette section, nous donnons des généralités sur la classification en utilisant l'apprentissage profond. Un domaine qui attire la communauté des chercheurs cette dernière décennie. Il a montré ses capacités de prédiction et de généralisation des modèles. Nous nous intéressons à un axe restreint qui est la classification des images médicale (images d'endoscopie par capsule sans fil (Wireless Capsule Endoscopy WCE).

2.6.1 Motivation

Les réseaux neuronaux, en particulier les réseaux neuronaux convolutifs (Convolutional neural network CNN) sont l'une des principales technologies utilisées pour effectuer l'analyse des images numériques, y compris la classification des images. Le CNN est typiquement modélisé comme le cortex visuel des mammifères, il est donc appliqué avec succès aux tâches de reconnaissance de la vision [55, 56]. La classification des images WCEs est une tâche difficile, car plusieurs facteurs produisent des changements négatifs dans les images, comme les réflexions spéculaires (saturation de la lumière), la distribution inégale des pixels (vignettage), les zones floues et les zones sombres [57, 58]. Les CNNs ont la capacité d'apprendre des caractéristiques directement à partir de l'ensemble de données d'image d'entrée. Dans ce qui suit, nous définissons la structure générale d'un CNN.

2.6.2 Architecture d'un CNN

Les CNNs sont similaires aux réseaux de neurones artificiels (Artificial neural netwirk ANN) dans la mesure où ils sont constitués de couches de neurones qui s'optimisent automatiquement par apprentissage. Il a été démontré que les CNNs sont fortement discriminants pour la reconnaissance des formes dans le monde réel et qu'ils ont la capacité d'apprendre les structures globales et locales des images [57, 58]. Ils constituent l'un des algorithmes d'apprentissage les plus puissants pour reconnaître les informations des images et ont montré un succès remarquable dans la segmentation, la classification, la détection et la récupération des images, de façon pertinente [58]. Comme les ANNs standards, les CNNs sont des réseaux multicouches, sauf que ces couches ne sont pas de simples perceptrons [55, 59]. Les couches CNNs se répartissent en quatre catégories : [55, 56, 59, 60] : couches de convolution, couches d'activation, couches de pooling et couches fully-connected.

— Couches de convolution : les opérations de convolution sont effectuées à l'aide de noyaux (ou filtres) convolutifs [55]. L'extraction de caractéristiques utiles est effectuée par des opérations de convolution, à partir de points de données localement corrélés, tout en préservant la relation spatiale entre les pixels [58]. Le résultat de la convolution sert d'entrée à la couche d'activation.

— Couches d'activation : elles sont constituées d'une unité de traitement non linéaire, appelée fonction d'activation [55]. Cette unité non linéaire intègre la non-linéarité dans l'espace des caractéristiques et contribue à l'apprentissage des abstractions, elle permet ainsi d'apprendre les différences entre les images, au niveau sémantique. Le résultat de l'activation sert d'entrée à la couche de pooling.

— Couche de pooling : cette couche effectue un sous-échantillonnage, pour résumer les résultats et rendre l'entrée invariante aux distorsions géométriques [58, 59]. Le principe est de calculer une valeur de sortie v pour une grille nxn de la carte d'activation, où v est la valeur maximale (max-pooling) ou moyenne (average-pooling) de cette grille dans la carte d'activation [55].

— Couches fully-connected : habituellement, elles sont utilisées à la fin d'un réseau CNN destiné à effectuer une classification [58]. Cette couche sert à représenter de manière compacte le signal d'entrée (par exemple des images) [55].

— Autres couches : il existe d'autres couches de régulation que les couches de mappage mentionnées ci-dessus, telles que la couche de batch normalization et la couche du dropout. Ces couches sont incorporées pour optimiser les performances du CNN [55]. Batch normalization est incorporée pour résoudre les problèmes dus au décalage de covariance interne (un changement dans la distribution des valeurs des unités cachées) dans les vecteurs de caractéristiques [58]. Le dropout est incorporé pour traiter le problème de suradaptation, en introduisant une régularisation dans le réseau, pour améliorer la généralisation en sautant aléatoirement certaines unités ou connexions avec une certaine probabilité.

La figure 2.1 illustre la structure générale d'un CNN.

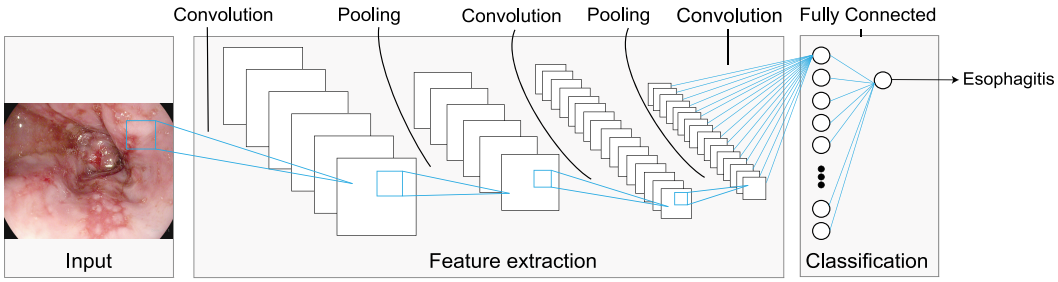

FIGURE 2.1 – Structure d'un réseau de neurones convolutifs CNN

2.6.3 Apprentissage par transfert

Pour former un CNN profond, il faut disposer d'un grand nombre de données d'apprentissage étiquetées, afin de garantir une performance de classification élevée. Néanmoins, l'acquisition et l'annotation d'images médicales est une tâche difficile. Lorsque de telles difficultés surviennent, l'utilisation d'attributs CNN pré-formé, tels que VGG 16 [61], via l'apprentissage par transfert s'est avérée efficace pour l'analyse d'images [61]. Le VGG-16 est un réseau pré-formé sur l'ensemble de données *ImageNet* [61], il possède une remarquable capacité d'extraction de caractéristiques qui implique un taux élevé de classification d'images. La structure générale d'un réseau neuronal convolutif VGG-16 est illustrée dans **la figure 2.2**.

Toutes les couches cachées sont équipées de la fonction ReLU (Rectified Linear Unit), qui calcule la fonction $f(x) = max(x, 0)$. Comme mentionné ci-dessus, la première couche convolutive reçoit une image de taille 224x224. Les couches convolutives utilisent un stride de 1 pixel. Les couches de max-pooling utilisent un stride de 2 pixels (pour le pooling spatial). Trois couches entièrement connectées suivent l'ensemble des couches convolutives. La taille de leurs canaux est respectivement de 4096, 4096 et 1000. Enfin, une couche softmax pour effectuer la classification. Le VGG-16 est utilisé de manière efficace, en raison de sa capacité de généralisation à d'autres ensemble de données [61].

2.6.4 Proposition d'un modèle de classification basé sur VGG-16

Dans notre proposition [10], seules les trois dernières couches du VGG-16 ont fait l'objet d'un réglage fin dans notre modèle. Cette décision a été motivée par les observations ci-dessous [62] :

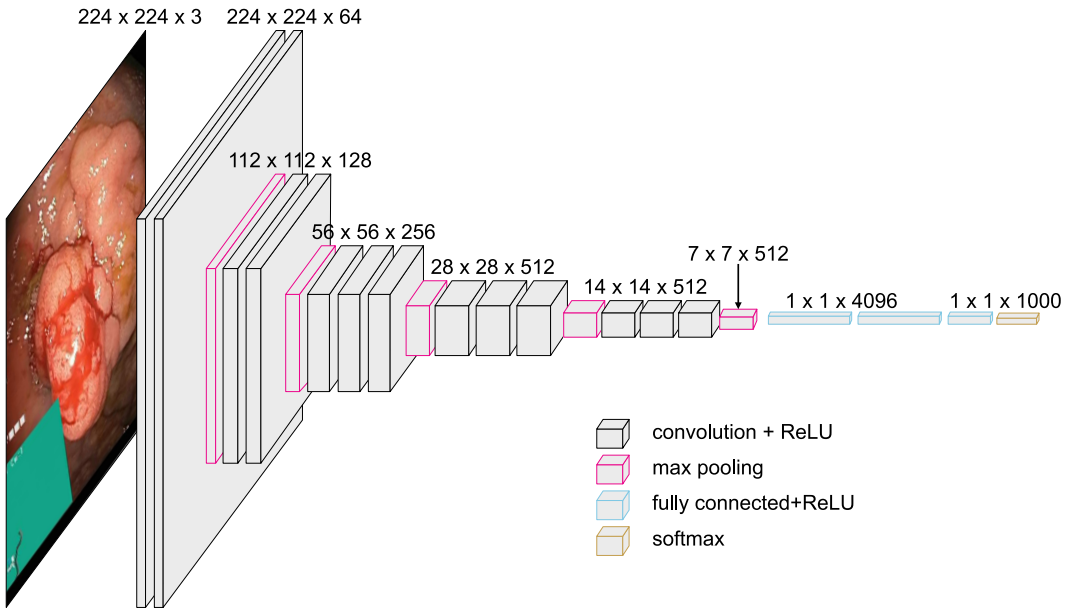

FIGURE 2.2 – La structure un model VGG-16

— Dans un CNN pré-formé, les premières couches contiennent des informations relatives aux bords et aux couleurs.

— Les couches suivantes contiennent des informations liées aux détails des classes.

L'apprentissage du VGG-16 a été effectué sur plus d'un million d'images et a été destiné à les catégoriser en 1000 classes [61]. Les trois dernières couches du VGG-16 effectuent la classification, en considérant ces 1000 classes. Par conséquent, ce sont ces couches qui doivent être affinées pour une nouvelle tâche de catégorisation [62, 63]. Le concept consiste donc à prendre toutes les couches telles qu'elles sont, à l'exception des trois dernières couches. Ensuite, on configure trois nouvelles couches, à savoir une couche entièrement connectée, une couche softmax, et une couche de sortie de classification, pour effectuer une nouvelle tâche de classification. La taille de la couche entièrement connectée doit être similaire au nombre de classes à distinguer [63]. La valeur de la taille dans notre travail actuel est de 8, ce qui correspond au nombre de classes. L'ensemble de donnée utilisé dans cette contribution est décrit dans une section ultérieure

Dans l'étape d'apprentissage, nous avons utilisé la fonction d'optimisation *RMSprop*, le momentum était de $0.9n$ le taux d'apprentissage initial était de $1e-1$ et le taux d'apprentissage final était de $1e-4$. Ceci afin d'obtenir la meilleure précision et de minimiser la fonction de

perte d'entropie croisée binaire, qui est donnée par E :

$$E = -\frac{1}{M}\sum_{m=1}^{M}[y_m \times \log(h(x_m) + (1-y_m) \times \log(1-h(x_m))] \tag{2.1}$$

Dans une deuxième série d'expériences, l'augmentation de l'ensemble de données a été effectuée pour résoudre le problème de surajustement (overfitting) dû à la petite taille de l'ensemble d'apprentissage. Cela a été réalisé en reflétant les images dans les axes horizontal et vertical de manière aléatoire. En effet, l'avantage de la réflexion est d'aider à l'apprentissage de la structure anatomique vue de différents angles.

2.6.5 Motivation de choisir le réseau VGG-16

Le choix du réseau pré-formé VGG-16, par rapport aux réseaux pré-formés AlexNet et InceptionNet est motivé par :

— VGG-16 est plus profond qu'AlexNnet, ce qui lui permet de mieux apprendre les caractéristiques.

— VGG-16 est plus simple qu'InceptionNet, ce qui lui permet une meilleure capacité de généralisation.

2.6.6 Expérimentation et résultats

Dans cette section nous allons présenté l'environement d'exécution et les résultats des expérimentations menées, ainsi que l'ensemble de données utilisé.

2.6.6.1 Détails d'implémentation et d'exécution

L'algorithme complet est implémenté en Python 3, en utilisant l'IDE Pycharm. Les bibliothèques Python les plus importantes utilisées dans l'implémentation sont : tensorflow utilisant keras et open CV. Le programme est exécuté sur un système équipé d'un processeur Intel Core i5-93000H, 2,40 GHz x 8, 16 Go de RAM avec une partition swap, et d'un GPU GeForce GTX 1650/PCIe/SSE2.

2.6.6.2 Ensemble de données endoscopiques : Kvasir

L'ensemble de données Kvasir contient des images de l'intérieur du tractus gastro-intestinal (GI), recueillies à l'aide d'un équipement endoscopique au "Vestre Viken Health Trust" en Norvège. Cet ensemble de données a été rendu accessible à l'automne 2017 dans le cadre d'un challenge "Medical Multimedia Challenge" de MediaEval, un programme d'évaluation comparative qui attribue des défis à la communauté scientifique. La résolution des images varie de $720x576$ jusqu'à $1920x1072$ pixels. Trois repères anatomiques majeurs, trois découvertes cliniquement significatives et deux procédures endoscopiques sont représentés dans les images. Les

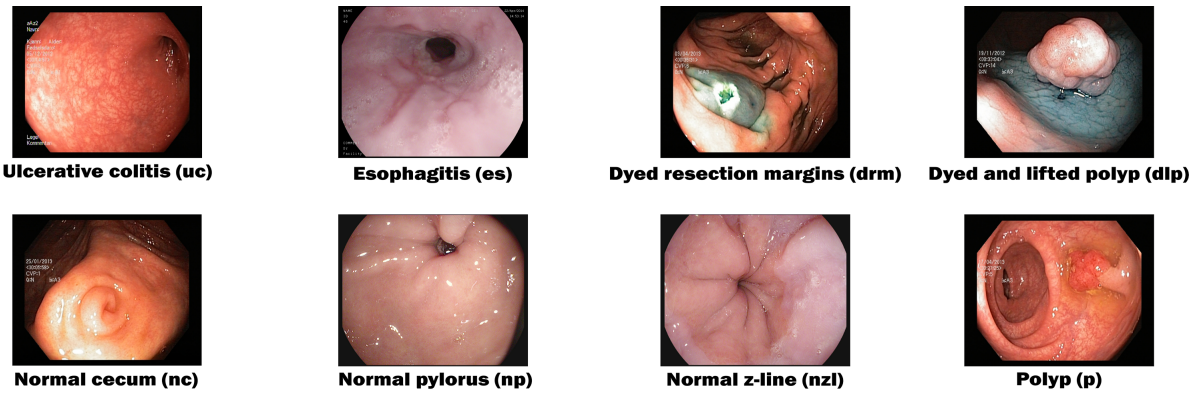

FIGURE 2.3 – Une image de chaque classe de l'ensemble Kvasir

images ont été annotées et vérifiées par des spécialistes. Sur certaines des images, le quart le plus à gauche de l'image est consacré aux annotations et aux informations sur l'image, tandis que la perspective anatomique occupe les trois quarts restants. Un encadré vert dans le coin inférieur gauche de plusieurs photos indique la position de l'endoscope dans le tube digestif. L'angle de vue, la taille, la luminance, le zoom et le point central sont tous différents.

Dans la **figure 2.3**, une image de chaque classe de la collection est choisie au hasard.

Repère anatomique : un repère anatomique est une caractéristique aberrante du système gastro-intestinal qui peut être vue à l'aide d'une procédure d'endoscopie. Il peut également s'agir de sites pathologiques typiques tels que des ulcères ou des inflammations. Les repères anatomiques inclus dans la collection Kvasit sont : *ligne Z*, *pylorus* et *cecum*.

Découvertes cliniques : l'endoscopie permet d'observer une lésion ou une altération de la muqueuse normale. Cette découverte peut être le symptôme d'une maladie en développement ou un précurseur. Les pathologies *ulcerative colitis polyps*, et *esophagitis* sont incluses dans l'ensemble de données Kvasir.

Procédures endoscopiques : deux séries d'images liées à l'ablation de lésions sont fournies, à savoir : *dyed and lifted polyp* et *dyed resection marginss*.

2.6.6.3 Résultat

Dans cette section, nous présentons les résultats de classification sur l'ensemble de données Kvasir. Ainsi, nous illustrons la comparaison de notre modèle proposé avec les résultats correspondant au défi *MediaEval* sur Kvasir. Avant de procéder à l'augmentation des données *Srie*1, notre modèle a atteint une accuracy de 96.9%. Après l'augmentation de données *Srie*2 notre modèle a atteint une accuracy de 98.8%. La **figure 2.4** montre la matrice de confusion de *Srie*2 comparant les performances interclasses. **Le tableau 2.2** montre que les résultats de nos propositions sont compétitifs avec les résultats du défi de *MediaEval* [10]. Cependant, cette comparaison n'est pas très juste, car les ensembles d'apprentissage et de test utilisés ne sont pas identiques. Néanmoins, les modèles proposés semblent performants sur chaque métrique fournie.

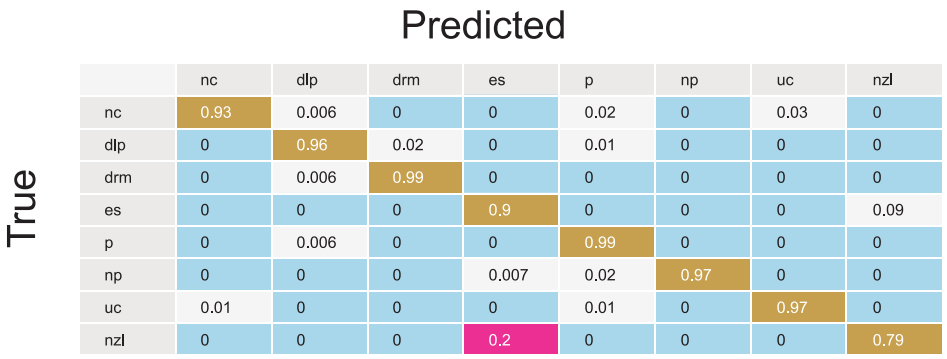

FIGURE 2.4 – Matrice de confusion démontrant la précision interclasse

TABLE 2.2 – Performance de notre model et les modèles du challenge MediaEval

Auteur	**Année**	**Acc**	**Pr**	**R**	**F1**
Series 1	2021	96.9	84.9	84.8	83.2
Series 2	2021	**98.8**	**87.3**	**88.6**	**87.3**
Agrawal et al [64]	2017	96.1	84.7	85.2	84.7
Pogorelov et al. [65]	2017	95.7	82.9	82.9	82.6
Naqvi et al. [66]	2017	94.2	76.7	77.2	76.7
Petscharnig et al. [67]	2017	93.9	75.5	75.5	75.5
Liu et al. [68]	2017	92.6	70.3	70.3	70.3

Après l'analyse de la matrice de confusion, nous observons que notre proposition est très performante pour les polypes et les marges de résection teintées. Nous pouvons également observer que la plus grande confusion se situe entre les classes Ligne-Z normale et Œsophagite. Nous pouvons justifier cela par l'observation faite par Pogorelov et al. [65]. En effet, les deux classes sont capturées à partir d'un emplacement anatomique similaire, sauf que l'œsophagite appartient à une zone malsaine et la z-line normale à une zone saine. De plus, **Tabel 2.2** montre que notre solution surpasse les résultats du défi de MediaEval en termes d'exactitude.

2.7 Conclusion

Dans ce chapitre, nous avons défini brièvement les modèles de classification les plus utilisés dans la littérature, à savoir les arbres de décisions DT, machines à vecteurs de support SVM et K-plus proches voisins KNN. Ensuite, nous avons décrit les métriques d'évaluation d'un modèle de classification à savoir l'accuracy Acc, la précision Pr, le rappel Rec, l'erreur Err, F_measure FM, la sensibilité Sn, la spécificité Sp, la moyenne géométrique GM et kappa Ka. Nous avons aussi défini quelques tests statistiques non paramétriques, pour la comparaison des modèles de classification, à savoir le test de $Friedman$, le test de $Nemenyi$ et le test de $Bonferroni - Dunn$. Dans une deuxième partie, nous avons donné un aperçu sur la classification des images

en utilisant l'apprentissage profond. Nous avons fini notre étude par la présentation d'un modèle en DL. Nous avons utilisé dans notre modèle la stratégie de l'apprentissage par transfert. Nous avons fait le réglage des trois dernières couches du CNN pré-formé VGG-16. Le modèle a été appliqué essentiellement pour catégoriser des images endoscopiques en 8 différentes classes. Le modèle a atteint une accuracy de 98.8% et a montré de meilleures performances par rapport aux modèles dévelopés dans le cadre du challenge MediaEval.

Chapitre 3

Sélection de caractéristiques pour la classification

3.1 Introduction

La majorité des problèmes de classification du monde réel nécessitent un apprentissage supervisé où les probabilités de classe sous-jacentes et les probabilités conditionnelles de classe sont inconnues et où chaque instance est associée à une étiquette de classe [16, 69]. Dans les situations réelles, les caractéristiques pertinentes sont souvent inconnues à priori [5]. Par conséquent, de nombreuses caractéristiques candidates sont introduites pour mieux représenter le problème. Malheureusement, la plupart des caractéristiques sont non pertinentes ou sont redondantes par rapport au concept cible [1, 5]. Dans de nombreuses applications du monde réel, la taille de l'ensemble des caractéristiques est tellement importante que l'apprentissage ne serait pas efficace. Le processus de la sélection des caractéristiques permet non seulement l'apprentissage du modèle de la classification plus rapidement, mais aussi de réduire la complexité du modèle, de le rendre plus facile à comprendre et d'améliorer les performances en termes d'exactitude et de précision [1, 16, 69]. Par conséquent, la sélection des caractéristiques permet la généralisation du modèle de classification et permet aussi d'éviter le surapprentissage du modèle et d'éviter la malédiction de la dimensionnalité [5]. Dans ce chapitre nous passons en revue des méthodes de sélection de caractéristiques pour la classification, nous classons ces méthodes par domaine d'applicaiton.

3.2 Recherche et indexation d'images

Les systèmes d'indexation et de recherche d'images basés sur le contenu (CBIR) analysent souvent le contenu des images en utilisant les caractéristiques dites de bas niveau, telles que la couleur, la texture et la forme. Pour obtenir des performances de recherche sémantique nettement supérieures, les systèmes récents ont tendance à combiner les caractéristiques de bas niveau avec des caractéristiques de haut niveau qui contiennent des informations perceptives

pour l'homme. Cependant, ces combinaisons augmentent les besoins en temps et en mémoire, ainsi que la complexité du processus d'extraction des caractéristiques. L'optimisation des performances des processus d'indexation et de recherche joue un rôle important dans la fourniture de services CBIR efficaces pour les systèmes limités. La sélection des caractéristiques est l'un des principaux défis pour l'optimisation des systèmes CBIR. Il s'agit de sélectionner les caractéristiques les plus importantes et leurs combinaisons pour décrire et interroger les éléments de la base de données afin de réduire la complexité de la recherche (temps et calcul) tout en maintenant des performances de recherche élevées.

Guldogan et al. [70] ont proposé un nouveau système de sélection de caractéristiques pour la recherche d'images basée sur le contenu. Le système proposé vise à améliorer les résultats de la recherche sémantique d'images, à réduire la complexité du processus de recherche et à améliorer la convivialité globale du système pour les utilisateurs finaux des moteurs de recherche multimédia. Ce nouveau système est constitué de trois critères de sélection des caractéristiques et une méthode de décision. Deux nouveaux critères de sélection de caractéristiques basés sur les relations innercluster et intercluster ont été proposés. Chaque critère est basé sur différentes associations d'affinité caractéristique-données pour définir la meilleure caractéristique discriminante et représentative des données. Ensuite, une méthode basée sur le vote majoritaire est adaptée pour une sélection efficace des caractéristiques et des combinaisons de caractéristiques. Les critères proposés produisent des résultats pour chaque caractéristique qui sont introduites dans le vote majoritaire. Les performances des critères proposés sont évaluées sur une grande base de données d'images et un certain nombre de caractéristiques et sont comparées aux techniques concurrentes de la littérature. Les expériences montrent que le système de sélection de caractéristiques proposé améliore les résultats de performance sémantique dans les systèmes de recherche d'images. L'inconvénient de cette proposition est que la sélection n'est pas automatique.

Allani et al. [71] ont proposé une approche qui adapte la sélection des caractéristiques visuelles au contenu sémantique assurant la cohérence entre elles. Etant donné que les approches de recherche et récupération d'images basées sur le contenu ont été marquées par l'écart sémantique (incohérence) entre la perception de l'utilisateur et la description visuelle de l'image. Cette incohérence est souvent liée à l'utilisation de caractéristiques visuelles prédéfinies choisies aléatoirement et appliquées quel que soit le domaine d'application. La première étape de leur approche est la conception des ontologies descriptives visuelles et sémantiques. Ces ontologies sont ensuite explorées par des règles d'association visant à lier un descripteur sémantique (un concept) à un ensemble de traits visuels. Les collections d'entités obtenues sont sélectionnées en fonction des images de requête annotées. Différentes stratégies ont été expérimentées et leurs résultats ont montré une amélioration de la tâche de récupération basée sur des sélections dynamique de caractéristiques pertinentes de bas niveau guidée par le contenu sémantique de l'image de requête. Malgré l'originalité de cette étude, elle représente des inconvénients. La

sélection n'est pas automatique, comme le cas de la plupart des méthodes de sélection des caractéristiques. Un autre inconvénient est que l'ensemble de données de test est relativement petit et très spécifique.

Xiang et al. [72] ont fait une étude à deux niveaux, le premier c'est l'extraction des caractéristiques et le deuxième c'est la sélection des caractéristiques. Le processus d'extraction de caractéristiques est analysé et une nouvelle approche de sélection de caractéristiques des bords est proposée. Une approche révisée d'extraction de caractéristiques structurelles basées sur les bords est introduite. Un algorithme principal de sélection de caractéristiques est également proposé pour l'analyse de nouvelles caractéristiques et la sélection de caractéristiques. Les résultats de la PFA (Principle Feature Analysis) sont testés et comparés à l'ensemble de caractéristiques d'origine, aux sélections aléatoires, ainsi qu'à ceux de l'analyse en composantes principales et de l'analyse discriminante linéaire multivariée. Les expériences ont montré que les caractéristiques proposées fonctionnent mieux que le moment d'ondelettes pour la récupération d'images dans une base de données d'images du monde réel et les caractéristiques sélectionnées par l'algorithme proposé donne des résultats comparables à l'ensemble de caractéristiques d'origine et de meilleurs résultats que les ensembles aléatoires. Dans leurs tests, ils ont utilisé un ensemble de données de 17695 images annotées, pour résoudre un problème de classification binaire.

3.3 Analyse génomique

Avec l'évolution des techniques informatiques, la quantité de données génomiques a augmenté de façon exponentielle, à un rythme rapide, ce qui rend difficile l'utilisation de ces données dans le domaine médical, tel que des dizaines de milliers de gènes sont mesurés dans un essai typique de microarray et un profil protéomique de spectrométrie de masse pour comprendre la fonction d'un organisme, ainsi que le comportement, la dynamique et les caractéristiques des maladies [1, 12, 73]. Par conséquent un prétraitement est nécessaire afin d'utiliser ces données de façon efficace en termes de temps et de précision de solution. L'une des façons les plus utilisée pour gérer la dimensionnalité est la sélection des caractéristiques pertinentes.

Dans [74], Afshar et al. ont proposé une nouvelle méthode linéaire de sélection de caractéristiques. Elle se déroule en deux grandes étapes essentielles, la première consiste à éliminer les caractéristiques non pertinentes et la deuxième consiste à évaluer la corrélation entre les caractéristiques pertinentes en utilisant la théorie des perturbations, les caractéristiques redondantes. Une classification par SVM et forêts aléatoires RF ont été utilisés pour effectuer la prédiction. La méthode a démontré ses performances et sa robustesse sur un ensemble de données synthétiques et dix ensemble de données génomiques réelles, par rapport à $mRMRM$ et d'autre méthodes de référence.

Tsamardinos et al. [75] proposent un algorithme de sélection de caractéristiques dans des contextes de big data qui peut combiner des coefficients de régression logistique locaux à des modèles globaux. L'algorithme est testé à l'aide de l'ensemble de données Single Nucleotide

Polymorphisms, par rapport aux modèles de régression logistique globaux produits par Apache MLlib 1 et montre une meilleure performance en termes du nombre de caractéristiques sélectionnées et la performance prédictive.

D'autre part, l'analyse du microbiote intestinal en relation avec les maladies mentales, en particulier la schizophrénie fait l'objet de l'étude de [76], dans laquelle Shen et al. réalisent plusieurs expériences en utilisant l'algorithme de sélection de caractéristiques de Boruta suivi d'un classifieur de forêt aléatoire. Les données utilisées sont des données réelles sur de 64 patients schizophrènes et de 53 patients sains et sont disponible publiquement. Cet ensemble de données porte pluparticulièrement sur le séquençage de l'ARNr 16S. Les résultats obtenus montrent qu'il existe certaines différences dans le microbiote intestinal entre les patients atteints de schizophrénie et les patients sains. Ces résultats pourraient être utilisés pour développer un diagnostic de la schizophrénie basé sur le microbiote.

3.4 Détection d'intrusion

Les systemes d'information partagés sur les réseaux connaissent une grande évolution et deviennent indispensable dans plusieurs domaines du quotidien. Cependant, la facilité que cette évolution apporte au quotidien un coût qui est une vulnérabilité accrue aux attaques malveillantes. La sécurité du système est donc une question importante pour protéger les réseaux de communication contre les attaques intruses. L'un des moyens de protéger les réseaux de communication est la détection des intrusions. Un certain nombre d'études ont déjà montré l'efficacité de l'apprentissage automatique dans la détection d'intrusion sur un réseau. L'un des aspects essentiels du développement d'un système de détection utilisant l'apprentissage automatique est la sélection des caractéristiques.

Dans [77], les chercheurs ont proposé une méthode améliorée de sélection des caractéristiques basée sur l'algorithme génétique (AG), appelée sélection des caractéristiques basée sur l'AG (GbFS), afin d'augmenter la précision des classifieurs. Ceci pour résoudre le problème de la sélection des caractéristiques dans le domaine de la sécurité des réseaux et de la détection d'intrusions. Plus précisément ce travail présente le réglage des paramètres pour la sélection des caractéristiques basée sur l'AG ainsi qu'une nouvelle fonction objective qui attribue les valeurs de fitness aux individus de la population GA permettant de sélectionner les chromosomes qui représentaient l'ensemble de caractéristiques optimal .Leur méthode a été testée sur trois ensembles de données de référence sur le trafic réseau, à savoir CIRA-CIC-DOHBrw2020, UNSW-NB15 et Bot-IoT. Une comparaison est également effectuée avec quatre méthodes de sélection de caractéristiques standard, à savoir, l'élimination de caractéristiques récursive, le sélecteur de caractéristiques séquentiel, la sélection de caractéristiques basée sur la corrélation et selectKbest. Les résultats montrent que les accuracies s'améliorent dans la plupart des cas et plus significativement en utilisant GbFS en atteignant une accuracy maximale de $99,80\%$. La limite de cette approche est le temps d'exécution, car le GA est utilisé au cœur de la solution actuelle

qui est fondamentalement lente en raison des multiples itérations répétitives et des opérations de reproduction.

Dans [78], les chercheurs ont proposé un algorithme de sélection de caractéristiques basé sur la combinaison de trois mesures MMFSA, où chaque mesure estime différentes informations qualitatives dans les caractéristiques. MMFSA a été testé sur quatre bases de données de référence, à savoir KDD'99, NSLKDD, CDMC2012 et CDMC2013. Dix classifieurs ont été utilisé, à savoir CART, C4.5, KNN, NNge, OneR, PART, RIDOR, SVM, DT et CR. Les expériences menées montrent que MMFSA surpasse, en termes d'efficacité de classification, chacune des mesures qu'il combine lorsqu'elles sont utilisées individuellement et les autres algorithmes de sélection de caractéristiques comparés. En effet, MMFSA utilise plusieurs mesures de sélection de caractéristiques, ce qui aide à sélectionner les caractéristiques les plus pertinentes. De plus, le point fort de cet algorithme est que la sélection se fait de façon automatique, sans prédéfinir manuellement le nombre de caractéristiques à sélectionner. Cependant, ce modèle présente une limite, l'apprentissage nécessite le choix d'une valeur de paramètre aléatoire. La valeur optimale du paramètre a été déterminée en testant plusieurs seuils.

Firuz et d'autre chercheurs [79] ont analysé les méthodes existantes de sélection des caractéristiques afin d'identifier les éléments clés des données de trafic réseau qui permettent la détection des intrusions. De plus, ils ont proposé une nouvelle méthode de sélection des caractéristiques qui relève le défi de considérer des caractéristiques d'entrée continues et des valeurs cibles discrètes. La méthode proposée donne de bons résultats par rapport aux méthodes de sélection de référence. Leurs résultats ont été utilisé pour développer un système de détection hautement efficace basé sur l'apprentissage automatique qui atteint une précision de $99,9\%$ pour distinguer les signaux DDoS des signaux bénins. Dans la phase de la sélection, ils ont utilisé trois méthodes de sélection standard : les algorithmes univariés basés sur la corrélation, les algorithmes univariés basés sur l'information mutuelle et les algorithmes de recherche avancée basés sur la corrélation. Ensuite, le nombre souhaité de caractéristiques est sélectionné en fonction du classement. On conclut que cette méthode sélectionne itérativement les caractéristiques en fonction de leur pertinence maximale et de leur redondance minimale. Les tests ont été faits sur l'ensemble de données CSE-CICIDS2018 qui se compose de 10000 instances. Les résultats obtenus ont amélioré la compréhension des caractéristiques pertinentes dans les données de trafic réseau et la construction des systèmes de détection d'intrusions puissants. La méthode proposée est très simple à mettre en œuvre et donne de bons résultats de détection d'intrusion sur des données réelles. Son originalité est qu'elle vise à résoudre la divergence entre les caractéristiques d'entrée continues (longueur des paquets, octets de sous-flux, etc.) et la variable cible discrète (bénigne/malveillante). Nous pensons que ces résultats peuvent être utiles aux experts qui s'intéressent à la conception et à la construction de systèmes automatisés de détection des intrusions. Bien que les avantages soient nombreux et motivants, sa limite est la même que la plupart des méthodes de sélection : la sélection n'est pas automatique, c'est à dire il faut au préalable introduire le nombre de caractéristiques à sélectionner.

3.5 Télédetection

La télédétection à partir des images de haute résolution est largement utilisée dans différents domaines, tels que la cartographie, la surveillance de la couverture terrestre, l'analyse de classification, la détection de routes et l'extraction automatique de bâtiments dans un environnement complexe. La méthode d'extraction de l'information la plus utilisée est la méthode basée sur les pixels ou objets. Les approches basées sur les pixels utilisent les pixels comme unités d'analyse de base et les approches basées sur les objets divisent une image en régions homogènes qui sont dit objets de différentes tailles contenant plusieurs pixels. Avec l'augmentation de la résolution de l'espace d'image, ces méthodes ne peuvent pas satisfaire le besoin d'extraction d'informations. Dans les deux cas, le grand nombre de caractéristiques générées entraîne une dégradation des performances. Par conséquent, la sélection de caractéristiques peut être appliquées pour résoudre ces problèmes.

Dans [80], les chercheurs ont traité un problème de télédétection pour les prévisions météorologiques. Afin de pallier aux problème de la grande dimensionnalité de l'espace de caractéristiques des données, ils ont proposé la technique "Tanimoto Correlation based Combinatorial MAP Expected Clustering and Linear Program Boosting Classification (TCCMECLPBC)". La première étape consiste à recueillir les données et les caractéristiques dans une grande base de données météorologiques. Ensuite, le coefficient de corrélation de Tanimoto est utilisé pour trouver la similarité entre les caractéristiques afin de sélectionner les plus pertinentes, cette étape a pour but d'éliminer les caractéristiques redondantes. Après avoir sélectionné les caractéristiques pertinentes, le processus de clustering attendu par MAP est effectué pour regrouper les données météorologiques, pour former des clusters. Dans ce processus, un certain nombre de clusters et de centroïdes de clusters sont initialisés. Ce processus de clustering comprend deux étapes, à savoir l'espérance (E) et la maximisation (M), pour découvrir la probabilité maximale de regroupement des données dans le cluster. Ensuite, le résultat du clustering est donné au classifieur de boosting par programme linéaire pour améliorer la performance de prédiction. La classification a été effectué en utilisant SVM et RF, sur deux ensembles de données météorologique, à savoir Hurricanes and Typhoons. Les résultats montrent que la technique TC-CMECLPBC améliore l'accuracy de la prédiction en moins de temps et avec un taux de faux positifs inférieur à celui des méthodes conventionnelles.

Shen et al. [81] ont comparé dix méthodes de sélection de caractéristiques appartenant à cinq groupes (méthodes basées sur la similarité, sur les statistiques, sur l'apprentissage clairsemé, sur la théorie de l'information et les wrappers) en fonction du F1 score et de la taille des données pour cartographier un paysage infesté par la mauvaise herbe Parthenium (Parthenium hysterophorus). Dans l'ensemble, les résultats ont montré que ReliefF (une approche basée sur la similarité) était la méthode de sélection des caractéristiques la plus performante, comme le démontrent les valeurs élevées du F1 score de la mauvaise herbe Parthenium et la petite taille des caractéristiques optimales sélectionnées. Bien que svm-b (une méthode wrapper) ait donné

les meilleurs résultats en termes d'exactitude, la taille du sous-ensemble optimal de caractéristiques sélectionnées était assez grande. Les résultats ont également montré que la taille des données affecte les performances des algorithmes de sélection des caractéristiques, à l'exception des méthodes basées sur les statistiques telles que l'indice de Gini-index, le F1 score et svm-b. Les résultats de cette étude fournissent une orientation sur l'application des méthodes de sélection des caractéristiques pour la cartographie précise des espèces végétales envahissantes en général et de la mauvaise herbe Parthenium, en particulier, en utilisant la nouvelle imagerie multispectrale à haute résolution temporelle.

Dans [82], les chercheurs ont proposé une nouvelle approche de sélection des caractéristiques, dans laquelle la méthode ReliefF présentée dans le chapitre 1, un algorithme génétique et la machine à vecteurs de support sont intégrés. L'algorithme ReliefF a été adopté pour filtrer de manière préliminaire les caractéristiques de haute dimension dans la base de données des caractéristiques. Après avoir éliminé les caractéristiques triées, le sous-ensemble de caractéristiques et les paramètres de la machine à vecteurs de support sont encodés dans le chromosome de l'algorithme génétique. Une fonction de fitness est construite en considérant la précision d'identification de l'échantillon, le nombre de caractéristiques sélectionnées et le coût de la caractéristique. La méthode proposée a été appliquée à des images à haute résolution obtenues à partir de différents capteurs, GF-2, BJ-2 et de véhicules aériens sans pilote (drones). Pour évaluer la précision, la matrice de confusion, l'exactitude, le rappel et le score F1 ont été appliqués. Les résultats montrent que la méthode proposée a permis de réduire les caractéristiques et que la précision globale était supérieure à 85%, avec des valeurs de coefficient de Kappa de $0,80$, $0,83$ et $0,85$, respectivement. L'efficacité temporelle de la méthode proposée était deux fois supérieure à celle du SVM avec toutes les caractéristiques. Les avantages de cette méthode est une grande réduction automatique des caractéristique et un temps de traitement réduit.

3.6 Classification du text

Le volume massif de données textuelles en ligne sur internet, comme les courriels, les sites sociaux et les bibliothèques, est d'une grande croissance. Par conséquent, la classification automatique du texte est devenue un axe de recherche très important pour rendre l'information textuelle utile [1, 12]. Cela peut traiter plusieurs problèmes du monde réel, tel que le filtrage du spam, l'analyse des sentiments et la classification des nouvelles. Les textes sont généralement représentés par une matrice document-terme de haute dimension et peu dense dans un espace ayant la dimensionnalité de la taille du vocabulaire, contenant les comptes de fréquence des mots [37]. La haute dimensionnalité peut causer certains problèmes, tels que la malédiction de la dimensionnalité et le surajustement du modèle de classification. Par conséquent, la sélection de caractéristiques peut être utilisée pour réduire la dimensionnalité afin d'augmenter la précision de la classification.

Dans [83], Thirumoorthy et al. ont proposé un schéma de sélection de caractéristiques basé sur la mesure de la distribution de la fréquence des termes. La classification a été effectuée en utilisant les modèles naïve de bayes et SVM sur deux ensembles de données de référence, à savoir WebKB et BBC. Les résultats de classification ont montré que leur schéma surpasse les méthodes de filtrage compétitives.

Dans [84], Liu et al. ont proposé une méthode de sélection de caractéristiques pour la classification de textes basée sur la recherche d'espaces de caractéristiques indépendants. Tout d'abord, une méthode de différence de fréquence relative entre les documents et les termes (RDTFD) est proposée pour diviser les caractéristiques de tous les documents textuels en deux ensembles de caractéristiques indépendants selon la capacité des caractéristiques à distinguer les échantillons positifs et négatifs, ce qui a deux fonctions importantes : l'une est d'améliorer la corrélation de classe élevée des caractéristiques et de réduire la corrélation entre les caractéristiques et l'autre est de réduire la plage de recherche de l'espace des caractéristiques et de maintenir une redondance appropriée des caractéristiques. Ensuite, la stratégie de recherche de caractéristiques est utilisée pour rechercher le sous-ensemble optimal de caractéristiques dans un espace de caractéristiques indépendant, ce qui peut améliorer les performances de la classification de textes. Ensuite, ils ont évalué plusieurs expériences sur six ensemble de données de référence. Les résultats expérimentaux montrent que la méthode RDTFD basée sur la recherche dans un espace de caractéristiques indépendant est plus robuste que les autres méthodes de sélection de caractéristiques.

Le problème des petits échantillons et de la grande dimensionnalité pour la classification des textes rend l'évaluation des méthodes de sélection des caractéristiques difficile car elle implique plusieurs critères. Par conséquent, la mise en œuvre d'une meilleure méthode d'évaluation prenant en compte plusieurs critères est nécessaire. Pour résoudre ce problème, une étude a été faite par Gang et al. [85]. Ils ont utilisé une méthode d'évaluation basée sur la prise de décision multi-critères MCDM (multiple criteria decision-making) pour évaluer les performances des méthodes de sélection de caractéristiques pour la classification de texte sur des ensembles de données avec un petit nombre d'échantillons. Après avoir obtenu des caractéristiques et des résultats de classification de texte à partir de 10 méthodes de sélection de caractéristiques communes et de trois classifieurs, les méthodes de sélection ont été évaluées en fonction des performances, de la stabilité et de l'efficacité de la classification. Ensuite, cinq méthodes MCDM ont classé les méthodes de sélection des caractéristiques en considérant toutes les mesures. Les chercheurs ont validé l'effet des cinq méthodes MCDM avec une expérience combinant 10 méthodes de sélection de caractéristiques, 9 critères de performance pour la classification binaire, 7 critères de performance pour la classification multi-classes, 5 méthodes MCDM et 10 ensembles de données de classification de texte. Les résultats montrent qu'aucune méthode de sélection de caractéristiques n'a atteint les meilleures performances sur tous les critères, quel que soit le nombre de caractéristiques et le classificateur choisi. Par conséquent, ils ont utilisé plus d'une mesure de performance pour évaluer les méthodes de sélection des caractéristiques. À partir des différents

résultats, ils ont fourni une recommandation de méthodes de sélection de caractéristiques, la méthode globalement préférée est DF (document frequency). Les résultats des tests trouvent que PROMOTHEE est le MCDM le plus adapté pour évaluer les performances du classifieur dans le cadre de la classification du texte. La limite de ce travail est que les expériences n'ont testé que 10 ensembles de données. De plus, il existe de nombreuses autres méthodes MCDM qui n'ont pas été analysées.

3.7 Méthodes de référence

Cette section est consacrée à la présentation des méthodes de sélection de caractéristiques qui seront par la suite utilisées pour la validation de notre approche proposée.

3.7.1 Infinite feature selection : a graph-based feature filtering approach $IFSS\_FS$

Dans [86], les chercheurs ont proposé une nouvelle méthode sélection de caractéristiques filtrantes $InfFS_S$ qui considère les sous-ensembles de caractéristiques comme des chemins dans un graphe, où un nœud est une caractéristique et une arête indique des relations par paires entre les caractéristiques, en traitant les principes de pertinence et de redondance. Les valeurs des chemins (sous-ensembles) sont évaluées en exploitant les propriétés des séries de puissance des matrices et en s'appuyant sur les principes fondamentaux des chaînes de Markov. La modélisation du graphe est comme suit :
Soit $G = (V,E)$ un graphe non dirigé et entièrement connecté, avec un ensemble de nœuds $V = v_1,...,v_n$ représentant un ensemble de n distributions de caractéristiques $F = f1,...,fn$ et un ensemble d'arêtes E modélisant les relations entre les paires de nœuds.

Le graphe G est représenté par sa matrice d'adjacence A , où chacun de ses éléments $A(i, j)$, $1 \leq i, j \leq n$, modélise la confiance que les caractéristiques fi et fj (les nœuds $\overrightarrow{vi}$ et $\overrightarrow{vj}$) sont toutes les deux bonnes candidates à sélectionner, grâce à une fonction de poids associée $\varphi(,)$:

$$A(i, j) = \varphi(\overrightarrow{vi}, \overrightarrow{vj})$$

Tel que, $\varphi(,)$ est une fonction positive à valeur réelle définissant la valeur de chaque arête. Elle additionne l'information de classe en utilisant le critère de Fisher et l'information mutuelle [87]. Ensuite, le problème de la sélection des caractéristiques a été considéré comme un problème de régularisation, où les caractéristiques sont des nœuds dans un graphe pondéré entièrement connecté et une sélection de l'ensemble des caractéristiques est un chemin de longueur l à travers les nœuds du graphe. Dans cette optique, la technique Inf_FS associe chaque caractéristique à un score provenant de fonctions par paire (les poids des bords) qui mesurent la pertinence et la non redondance. Ce score peut s'expliquer de différentes manières : sous l'angle d'une série de matrices, il indique la valeur qu'une caractéristique peut apporter dans une sélection éventuellement infinie de caractéristiques. Alternativement, dans une perspective de chaîne de

Markov absorbante, le score indique combien de fois une caractéristique serait associée aux autres indices comme complémentaires, avant de terminer le processus de sélection. Un sous-ensemble précis de caractéristiques peut être fourni, en examinant la distribution de ces scores. La méthode a été testé sur cinq ensembles de données génomique à deux classes. Les résultats ont surpassé les résultats de 18 approches compétitives.

3.7.2 Simple strategies for semi-supervised feature selection

Konstantinos et al. [88] ont testé empiriquement deux stratégies simples indépendantes du classificateur. À partir des données binaires étiquetées et d'autres non étiquetées, ils ont supposé que les données non étiquetées sont toutes positives (ou toutes négatives). Ces approches minimalistes, apparemment naïves, n'ont pas encore été étudiées en profondeur, avant l'étude menée par Konstantinos et al. Grâce à des études théoriques et empiriques, ils ont montré qu'elles fournissent des résultats puissants pour la sélection de caractéristiques, via des tests d'hypothèse et le classement de caractéristiques.
Dans leurs tests empiriques, ils ont utilisé deux méthodes proposées par d'autres chercheurs, à savoir *IAMB* et *JMI*. EN utilisant *IAMB*, dans la phase de test d'hypothèse de la découverte de la couverture de Markov, ils ont utilisé des nœuds semi-supervisés dans le réseau bayésien.

3.8 Conclusion

Dans ce chapitre, nous avons présenté quelques travaux de la littérature qui traitent le problème de la sélection des caractéristiques. Nous avons classé les travaux par domaine d'application, à savoir : la recherche et indexation d'images, l'analyse génomique, la détection d'intrusion, la télédétection, l'exploration du texte. De plus, nous avons décrit en détail deux méthodes de sélection avec lesquelles nous avons comparé l'approche proposée dans le chapitre 4, à savoir : Infinite feature selection : a graph-based feature filtering approach $IFSS\_FS$ et Simple strategies for semi-supervised feature selection $SSSFS$. Nous concluons que bien que les efforts de la communauté de chercheurs fassent et bien que la sélection de caractéristiques ait été appliquée avec succès dans différents scénarios dans lesquels des ensembles de données de haute dimension sont présents, tels que l'analyse génomique, la classification des images et la classification des textes...etc, les travaux présentes toujours des limites et le problème de sélection de caractéristiques reste toujours un axe de recherche important.

Chapitre 4

Sélection des caractéristiques basée sur le calcul des valeurs propres avec contraintes

4.1 Introduction

Comme nous avons mentionné dans le chapitre 1, les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection sont classées en trois catégories principales [1, 12] : ”filter”, ”wrapper” et ”embedded”. Les méthodes de sélection des caractéristiques sont aussi basées sur le classement des caractéristiques, le clustering des caractéristiques ou l'optimisation [5].

L'étude des travaux connexes que nous avons présentées dans notre contribution [5] a montré les limites suivantes :

— Les approches basées sur le classement présentent un sens intuitif lors de la sélection des caractéristiques. Cependant, elles ne tiennent pas compte des relations implicites et explicites entre les caractéristiques.

— Les approches basées sur le clustering ont plusieurs limites telles que : (a) dans la plupart d'entre elles, l'utilisateur doit spécifier le nombre de clusters à l'avance, (b) elles sont très sensibles aux paramètres d'entrée.

— Les approches basées sur l'optimisation partent d'un sous-ensemble aléatoire de caractéristiques comme population initiale. Par conséquent, les résultats obtenus sont sensibles à cette initialisation et leur efficacité dépend fortement du choix initial.

Au cours de cette thèse, nous avons proposé une nouvelle approche de sélection de caractéristiques $Eigen\_FS$, en se basant sur les limites des travaux connexes mentionnées ci-dessus. Notre approche est une extension pour combler les lacunes et elle bénéficie des avantages des approches basées sur le classement et des approches basées sur l'optimisation. Elle est basée sur le calcul des valeurs propres avec des contraintes linéaires. $Eigen\_FS$ comprend les étapes suivantes :

— Formuler le problème de sélection comme un problème d'optimisation général,

— faire une approximation du problème à un problème de coupe maximale ;

— et le réduire en un problème de calcul de valeurs propres avec des contraintes linéaires.

Afin de résoudre notre problème, nous avons utilisé une approche itérative efficace basée sur la méthode d'itération de puissance pour le calcul des valeurs propres, à savoir "Projected power method" [89]. Cette méthode converge vers la solution optimale et est par conséquent appliquée à la résolution de divers problèmes d'optimisation. Les caractéristiques principales de notre contribution sont les suivantes :

— Elle considère la redondance entre les caractéristiques et la pertinence entre les caractéristiques et la classe cible.

— Le point de départ pour déterminer les meilleures caractéristiques pertinentes est l'ensemble des valeurs propres de la matrice de performances.

— Le problème formulé est résolu par une méthode itérative qui converge forcément vers une solution optimale.

Donc notre objectif principal est d'optimiser un critère prenant en compte à la fois la redondance et la pertinence. Cette nouvelle approche de sélection de caractéristiques basée sur l'optimisation est ensuite testée sur 20 ensembles de données UCI [90] avec différents types de caractéristiques. De plus, notre approche a été testé sur des données médicales réelles, cela va être présenté dans le Chapitre 6.

Nous commençons ce chapitre par introduire les notions de bases nécessaires pour la modélisation de notre approche. Ensuite nous présentons une nouvelle approche de sélection de caractéristiques basée sur le calcul des valeurs propre $Eigen\_FS$. Le protocole expérimental et les résultats seront présentés dans le chapitre suivant.

4.2 Contexte et motivation

Dans cette section, nous présentons les notions de bases qui constituent les étapes de notre formulation et notre solution.

4.2.1 Problèmes d'optimisation combinatoire

L'optimisation combinatoire est le processus de recherche des maxima (ou minima) d'une fonction objectif F dont le domaine est un espace discret mais large [91]. Considérons le problème d'optimisation combinatoire général suivant :
Soit $\mathbb{F}$ une famille de sous-ensembles d'un ensemble fini E et soit $w : \mathbf{E} \longrightarrow \mathbb{R}$ une fonction de poids à valeur réelle définie sur les éléments de $\mathbf{E}$. L'objectif du problème d'optimisation combinatoire est de trouver $F^* \in \mathbb{F}$, tel que [91] :

$$w(F^*) = \min_{F \in \mathbb{F}} w(F)$$

Où,

$$w(F) := \sum_{e \in \mathbb{F}} w(e)$$

Pour traduire le problème d'optimisation combinatoire en un problème d'optimisation dans $\mathbb{R}^E$, nous pouvons représenter chaque $F \in \mathbb{R}$ par son vecteur d'incidence.

$$\chi_e^F = \begin{cases} 1 & \text{si } e \in F \\ 0 & \text{sinon.} \end{cases}$$

Alors, soit $S = \{\chi_F : F \in \mathbb{F}\} \subseteq \{0,1\}^E$ l'ensemble des vecteurs d'incidence des ensembles dans $\mathbb{F}$, le problème d'optimisation correspondant est le suivant :

$$min\{w^T x : x \in S\}$$

4.2.2 Problème de coupe maximale "Max cut problem"

Dans cette section, nous allons présenter le problème de coupe maximale et coupe maximal normalisée. Nous allons ensuite faire une démonstration de l'approximation du problème de sélection de caractéristique au problème de coupe maximale.

4.2.2.1 Présentation du problème

Pour un graphe donné **G** , une coupe maximale est une coupe dont la taille est au moins égale à la taille de toute autre coupe [92, 93]. Autrement dit, il s'agit d'une partition des sommets du graphe en deux ensembles complémentaires **S** et **T**, de telle sorte que le nombre d'arêtes entre l'ensemble **S** et l'ensemble **T** soit le plus grand possible. Le problème consistant à trouver une coupe maximale dans un graphe est connu sous le nom de problème de la coupe maximale.

La coupe normalisée [89] : elle est définie sur un graphe non orienté pondéré $G = (V,E)$, où les nœuds du graphe sont les objets de données à regrouper et où il y a une arête entre chaque paire de nœuds. Le poids de chaque arête $w(u,v)$ est exprimé par une fonction de similarité positive entre les nœuds u et v. Dans le partitionnement à deux voies, un graphe $G = (V,E)$ est découpé en deux ensembles disjoints A et A' , $A \cup A' = V$, $A \cap A' =$, en supprimant les arêtes reliant les deux ensembles. Ce partitionnement peut être décrit par une fonction d'étiquetage y qui indique les deux ensembles disjoints.

$$y(u) = \begin{cases} 1 & \text{si } u \in A \\ -1 & \text{si } u \in A' \end{cases}$$

et la coupe de ce partitionnement peut être définie comme la somme des poids des arêtes qui relient les nœuds avec des étiquettes différentes :

$$cut(A,A') = \sum_{u \in A, v \in A'} w(u,v)$$

Intuitivement, on peut minimiser la valeur de coupe pour partitionner les données, ce qui est généralement appelé coupe minimale. Cependant, le critère de coupe minimale favorise la coupe de petits morceaux de nœuds isolés dans le graphe, ce qui est indésirable dans la plupart des cas.

4.2.2.2 Approximation du problème de sélection

Pour proposer une nouvelle méthode d'approximation, le problème combinatoire sous-jacent doit être étudié. Pour ce faire, nous pouvons formuler le problème de sélection de caractéristique comme suit :

$$\begin{aligned} &\max w^t M w; \\ &s.t \sum_{i=1}^{n} w_i = k \\ &w_i \in \{0,1\} \end{aligned} \tag{4.1}$$

Tel que :

— la taille du sous ensemble caractéristiques à sélectionner est k,

— M est la matrice de performance des caractéristiques.

par conséquent, la solution de ce problème revient à trouver l'ensemble de caractéristiques de taille k qui maximise la performance du sous-ensemble. Le problème de programmation (0,1)-quadratique (4.1) a attiré beaucoup d'études théoriques en raison de son importance dans les problèmes combinatoires. Ce problème peut simplement être transformé en un problème de programmation (-1,1)-quadratique :

$$\begin{aligned} &\max_z \frac{1}{4} z^t M z + \frac{1}{2} z^t M e + c; \\ &s.t \sum_{i=1}^{n} z_i = 2 \times k - n \\ &z_i \in \{-1,1\} \end{aligned} \tag{4.2}$$

Avec $z = 2w - e$, tel que e est un vecteur de 1. De plus c est une constante égale à $\frac{1}{4} e^t M e$ et peut être ignorée puisqu'elle ne dépend pas de z. Pour écrire la fonction objective dans (4.2) de façon homogène, nous définissons une nouvelle matrice A de taille $(n+1) \times (n+1)$ à partir de la matrice M en ajoutant une 0^{eme}ligne et une colonne [94].

$$A = \begin{bmatrix} 0 & e^T M \\ M^T e & M \end{bmatrix} \tag{4.3}$$

En ignorant le facteur constant $\frac{1}{4}$ dans (4.2), la forme homogène équivalente de (4.1) peut s'écrire [94] :

$$\begin{aligned} &\max z^t A z; \\ s.t &\sum_{i=1}^{n} z_i z_0 = 2k - n \\ &z_i \in \{-1, 1\} \end{aligned} \tag{4.4}$$

z est maintenant un vecteur de dimension $n+1$ avec le premier élément $z_0 = \pm 1$ comme variable de référence et qui sera ignorée. Le problème d'optimisation dans (4.4) peut être vu comme une instance du problème de coupe maximale [95] avec une contrainte de cardinalité supplémentaire.

4.2.3 Valeurs propre d'une matrice

Soit A une matrice nn, alors un vecteur x non nul dans R^n est appelé un vecteur propre de A (ou de l'opérateur matriciel TA) si Ax est un multiple scalaire de x ; c'est-à-dire, pour un certain scalaire λ [96] :

$$Ax = \lambda x$$

Le scalaire λ est appelé une valeur propre de A, et on dit que x est un vecteur propre de A correspondant à λ.

4.3 Prérequis mathématiques

Dans cette section, certaines formulations mathématiques du problème de sélection des caractéristiques sont présentées.

4.3.1 Mesures d'évaluation

Afin d'obtenir une plus grande pertinence, les caractéristiques ayant une *information mutuelle* plus élevée avec la classe cible sont sélectionnées. De même, afin d'éviter la redondance des caractéristiques, donc garantir une plus grande diversité, les caractéristiques ayant un plus grand score de *Fisher* entre elles seront sélectionnées. La pertinence et la redondance des caractéristiques sont stockées dans une matrice carrée symétrique, qu'on appelle *matrice de performance*. Dans ce qui suit, nous allons définir ces trois notions de base.

4.3.1.1 Information mutuelle

L'information mutuelle (IM) est une mesure de la quantité d'information entre deux variables aléatoires, elle est nulle si et seulement si les variables sont indépendantes [87]. La plupart des

études se concentrent sur la conception d'une procédure de sélection de caractéristiques efficace basée sur le concept de IM.

Information mutuelle [5, 87, 97] : *soient x et y deux variables aléatoires, leur information mutuelle $MI(x,y)$ est définie en fonction de leurs probabilités marginales $p(x), p(y)$ et leur fonction de masse conjointe $p(x,y)$:*

$$MI(x,y) = \int\int p(x,y) log \frac{(p(x,y))}{(p(x)p(y))} dxdy$$

$$MI(x,y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{(p(x,y))}{(p(x)p(y))} \tag{4.5}$$

4.3.1.2 Score de Fisher

Score de Fisher [97] : *le score de Fisher est calculé comme suit :*

$$Fisher'_i = \frac{|\mu_{i,1} - \mu_{i,2}|^2}{\sigma^2_{i,1} + \sigma^2_{i,2}} \tag{4.6}$$

où $\mu_{i,c}$ et $\sigma_{i,c}$ sont la moyenne et l'écart type, respectivement, supposés par la i^{me} caractéristique lorsque l'on considère les échantillons de la c^{me} classe, $1 \leq c \leq C$, tel que le problème de classification contient C classes. La généralisation multiclasse est donnée par :

$$Fisher_i = \sum_{c=1}^{C} \frac{(\mu_{i,g} - \hat{\mu}_i)^2}{E_i^2} \tag{4.7}$$

Oú

$$E_i^2 = \sum_{c=1}^{C} (\sigma_{i,c})^2 \tag{4.8}$$

où $\hat{\mu}_i$ et E_i désignent la moyenne et l'écart type de l'ensemble de données correspondant à la caractéristique f_i. Cela tient compte de la compacité intra-classe et de la séparation inter-classe induite par différentes caractéristiques. Les scores finaux sont normalisés pour avoir un maximum de 1 et un minimum de 0 . Plus $Fisher_i$ est proche de 1, moins la i^{me} caractéristique est redondante, puisque son domaine ne chevauche pas les autres.

4.3.2 Matrice de performance

Matrice de performance *une matrice de performance $M \in R^{nxn}$ est réglée pour enregistrer les performances de toutes les caractéristiques. Les entrées diagonales $M_{i,i}$ représentent la pertinence de la i^{me} caractéristique par rapport à la classe cible, et les entrées non diagonales $M_{i,j}$ mesurent la diversité par paire, entre la i^{me} et la j^{me} caractéristiques.*

4.4 Modélisation de l'approche

Nous commençons par le remplissage de la matrice la performance. Ensuite nous démontrons la formulation de notre problème, depuis un problème d'optimisation combinatoire à un problème de calcul de valeurs propres d'une matrice avec contraintes linéaire. Nous expliquons ensuite la méthode utilisée pour la résolution du problème. Nous finissons par adapter notre formulation finale à l'algorithme de la solution.

4.4.1 Configuration de la matrice

Soit $M \in R^{nxn}$ la matrice de performance, on a :

$$M_{i,j} = \begin{cases} MI(f_i, C) & \text{si } i = j \\ Fisher(f_i, f_j) & \text{sinon.} \end{cases} \tag{4.9}$$

Ainsi, une grande valeur de $M_{i,i}$ correspond aux caractéristiques les plus pertinentes et une grande valeur de $M_{i,j}$ correspond aux caractéristiques les moins redondantes. Par conséquent, une grande valeur de $\sum M_{i,j}$ implique un bon sous-ensemble.

4.4.2 Formulation d'un problème d'optimisation combinatoire

Sur la base de la configuration de la matrice de performance, nous formulons la sélection du sous-ensemble de taille k comme suit [5, 46] :

$$\begin{aligned} &\max w^t M w; \\ &s.t \sum_{i=1}^{n} w_i = k \\ &w_i \in \{0,1\} \end{aligned} \tag{4.10}$$

Où $w_i = 1$, the i^{me} caractéristique sera sélectionnée et sa pertinence sera prise en compte dans la fonction objectif. La contrainte $\sum_{i=1}^{n} w_i = k$ contrôle la taille du sous-ensemble sélectionné. Le problème (1) est un problème d'optimisation standard 0-1, qui est NP-Hard. Dans ce qui suit, nous le reformulons pour le résoudre efficacement.

4.4.3 Approximation max-coupe

Comme mentionné précédemment, nous considérerons chaque instance du problème de sélection de caractéristiques comme une instance du problème de max-coupe, ce choix est motivé par le fait que la méthode de puissance projetée est une méthode efficace proposée récemment pour résoudre le problème de max-coupe. En effet, la complexité de la "Méthode de la Puissance Projetée" est de $O(n^2)$ pour chaque boucle et il est prouvé après une expérimentation extensive qu'elle converge après une certaine itération [89].

Dans l'expression mathématique (1) du problème, les poids w_i appartiennent à l'ensemble $\{0,1\}$. Cependant, dans Mc-k $w_i \in \{-1,1\}$. Par conséquent, afin de réduire notre problème de sélection de caractéristiques à un problème de coupe maximale, nous effectuons la transformation de variable suivante.

Soit $z_i = 2w_i - 1$. Dans ce cas, quand $w_i \in \{0,1\}$, $z_i \in \{-1,1\}$. En suivant la démonstration de la section 4.2.2.2, le problème d'optimisation sera :

$$\begin{aligned} &\max z^t Az; \\ s.t &\sum_{i=1}^{n} z_i = 2k - n + 1 \\ &z \in \{-1,1\}^{(n+1)} \end{aligned} \tag{4.11}$$

Comme mentionné dans la même section, z est maintenant un vecteur de dimension $n+1$ avec le premier élément $z_0 = \pm 1$ qui sera ignorée. Le problème d'optimisation dans (4.11) est une instance du problème de coupe maximale [95] avec une contrainte de cardinalité supplémentaire.

4.5 Valeurs propres avec contraintes linéaires

Dans l'expression (4.11), les k premières meilleures valeurs dans $z_{(1:n)}$ indiquent que les caractéristiques correspondantes sont sélectionnées. La contrainte discrète $z_i \in \{-1,1\}$ rend notre problème NP-Hard. Ainsi, afin de relaxer ce problème, nous remplaçons la contrainte discrète $Z \in \{-1,1\}^{(n+1)}$ par la contrainte de norme $\|z\| = \sqrt{n+1}$ [98]. La formulation du problème devient alors :

$$\begin{aligned} &max(z^T Az) \\ s.t &z^T e = 2k - n + 1 \\ &\|z\| = \sqrt{n+1} \\ &z_0 = 1 \end{aligned} \tag{4.12}$$

Nous écrivons les contraintes ci-dessus sur le vecteur z, de manière compacte comme $Bz = c$, où :

$$B = \begin{bmatrix} 11.. & .1 \\ 10.. & .0 \end{bmatrix}$$
$$c = (2k - n + 1, 1)$$

La matrice B est construite pour satisfaire la première et la troisième contrainte de la fonction objective. La formulation du problème devient alors :

$$\begin{aligned} & maxz^T Az \\ & Bz = c \\ & \|z\| = \sqrt{n+1} \end{aligned} \tag{4.13}$$

La relaxation ultime (4.13) est non convexe et est connu pour donner la plus grande valeur propre de A [99], avec contraintes linéaires. Ce problème est largement étudié dans [100] où il est appelé un problème aux valeurs propres avec contraintes. Une nouvelle méthode numérique pour le problème aux valeurs propres avec contraintes (4.13) a été étudiée par *Xu, Li et Schuurmans* dans [89], c'est la méthode de la puissance projetée (Projected Power Method). La résolution de du problème (4.13) est étudiée dans la section suivante. Toutes les étapes de notre approche sont résumées dans la **Figure 4.1**.

4.6 Solution avec méthode de puissance projetée

L'idée principale de cette méthode est de combiner la méthode de puissance ("Projected Method") avec une étape de projection qui garantit que le vecteur résultant satisfait la contrainte linéaire. On commence par définir la matrice de projection : $P := I - B^T(BB^T)^{-1}B$ et le vecteur $n_0 = B^T(BB^T)^{-1}c$. Notons que pour $Bz = c$;

$$\begin{aligned} Pz &= [I - (B^T(BB^T)^1B)z] \\ &= z - B^T(BB^T)^1Bz \\ &= z - B^T(BB^T)^1c \\ &= z - n_0 \end{aligned} \tag{4.14}$$

Maintenant, nous utilisons ces définitions afin de transformer le problème d'optimisation. À partir de (4.14) on a :

$$\begin{aligned} \mathrm{n}_0 &= z - Pz \\ \|n_0\|^2 &= \|z\|^2 - 2z^T Pz + z^T Pz \\ \|n_0\|^2 &= 1 - z^T Pz \\ z^t PPz &= 1 - \|n_0\|^2 \\ \|zP\|^2 &= 1 - \|n_0\|^2 \end{aligned} \tag{4.15}$$

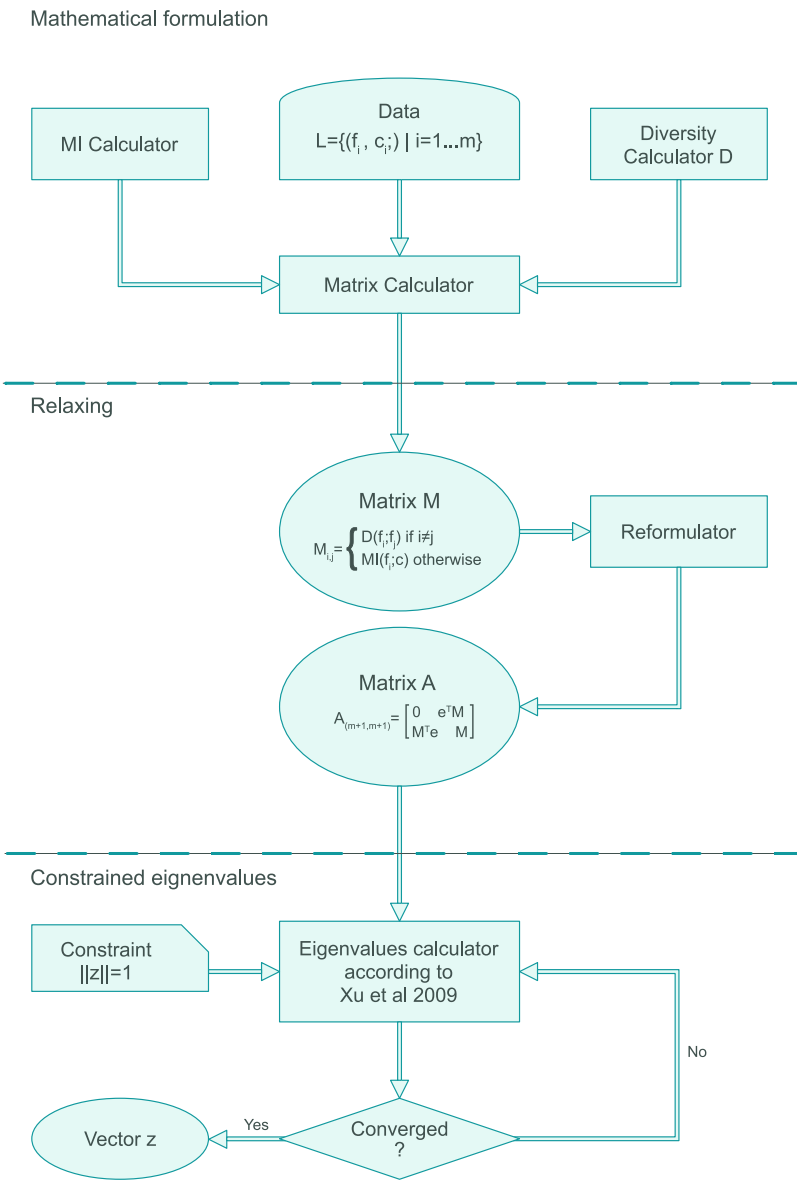

FIGURE 4.1 – Flowchart of the proposed method

Définissons $\gamma = \sqrt{1-||n_0||^2}$ et $b_0 := PAn_0$. Par conséquent :

$$\begin{aligned} \max_z z^t Az &= \max_z [(n_0+Pz)^T A(n_0+Pz)] \\ &= \max_z [n_0^T An_0 + 2z^T PAn_0 + z^T PAPz] \\ &= \max_z [n_0^T An_0 + z^T PAPz + 2z^T Pb_0] \end{aligned} \tag{4.16}$$

Nous remarquons que la valeur du premier terme est constante, il suffit donc de trouver le vecteur z qui résout :

$$\begin{aligned} \max_{z} [z^T PAPz + 2z^T Pb_0] \\ s.t \\ \|Pz\|^2 = \gamma^2 \end{aligned} \tag{4.17}$$

Le lagrangien du problème d'optimisation (4.17) est :

$$\mathcal{L}(\lambda, z) = \frac{1}{2} z^T PAPz + z^T Pb_0 - \frac{\lambda}{2}(\|Pz\|^2 - \gamma^2)$$

On prend maintenant les dérivées partielles de $\mathcal{L}$ par rapport à z et λ pour obtenir les équations suivantes :

$$\begin{aligned} (PA - \lambda I)Pz = -b_0; \\ \|Pz\|^2 = \gamma^2. \end{aligned} \tag{4.18}$$

La première équation de (4.18) doit être satisfaite pour atteindre le point critique du problème d'optimisation (4.17). À partir de la première équation de (4.17), la définition de b_0, et en supposant $\lambda > 0$, nous obtenons :

$$\begin{aligned} (PA - \lambda I)Pz &= -b_0 \\ PAPz - \lambda Pz &= -PAn_0 \\ PAPz + PAn_0 &= \lambda Pz \\ PA(Pz + n_0) &= \lambda Pz \\ PAz &= \lambda Pz \\ \|PAz\| &= \lambda \|Pz\| \\ \frac{\|PAz\|}{\|Pz\|} &= \lambda \end{aligned} \tag{4.19}$$

Par conséquent, $\lambda = \frac{\|PAz\|}{\gamma}$. Ceci est utilisé avec la première équation de (4.18) pour vérifier la convergence du vecteur z_{k+1} dans **l'algorithme 4** proposé dans [89].

La méthode de la puissance projetée [89] se comprend intuitivement à chaque étape où v_k est multiplié par la matrice A. Puis on le multiplie par la matrice de projection P. Cela va projeter le vecteur sur l'hyperplan $Bz = c$. Ce vecteur est ensuite redimensionné à la longueur γ et ajouté à n_0. Cela garantit que le vecteur résultant satisfait les deux contraintes. Les étapes de cet algorithme sont illustrées dans **la figure 4.2**. Cet algorithme est garanti de converger vers la solution globalement optimale du problème (4.13). Ce qui a été démontré dans la contribution originale [89].

Algorithm 4: FS via constrained Eigenvalues optimization

Input : Matrix M recording the features performance

-Calculate the matrix A, according to (4.3)

-Solve the optimization problem (8) :

H=AA^T

i=0 ;

P=I-B^T (BB^T) $^{-1}$ B ;

n_0 =B^T (BB^T) $^{-1}$ c ;

$\gamma = \sqrt{1-||n_0||^2}$;

$z_0 = \gamma \frac{PHn_0}{||PHn_0||} + n_0$;

Repeat

$u_{i+1} = \gamma \frac{PHz_i}{||PHz_i||}$;

$z_{i+1} = u_{i+1} + n_0$;

$i = i+1$;

Until z converges

Return z.

Le processus de l'algorithme 4 est illustré dans un graphique en 3 dimensions, dans **la figure 4.2**.

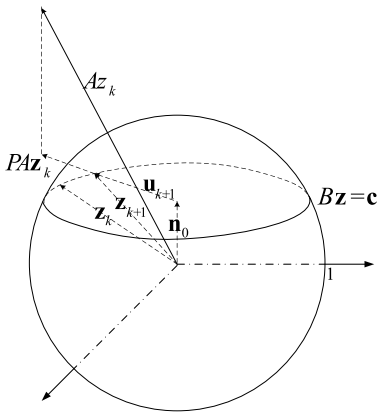

FIGURE 4.2 – Une vue géométrique de la méthode de puissance projetée

4.7 Adaptation de notre problème

La méthode de puissance projetée résout directement le problème d'optimisation suivant :

$$\begin{aligned} & maxz^T Az \\ & Bz = c \\ & \|z\| = 1 \end{aligned} \tag{4.20}$$

Alors que dans notre formule finale, nous avons :

$$\begin{aligned} & maxz^T Az \\ & Bz = c \\ & \|z\| = \sqrt{n+1} \end{aligned} \tag{4.21}$$

Ce qui affecte directement notre démonstration. Plus précisément, la ligne 3 de l'expression (4.15).

Pour pallier à ce problème d'incohérence, nous revenons à l'expression (4.11), qui devient :

$$\begin{aligned} & \max v^t Av; \\ & s.t \sum_{i=1}^{n} v_i = \frac{2k-n+1}{\sqrt{n+1}} \\ & v \in \{\frac{-1}{\sqrt{n+1}}, \frac{1}{\sqrt{n+1}}\}^{(n+1)} \end{aligned} \tag{4.22}$$

Par conséquent, en remplaçant la contrainte linéaire $\sum_{i=1}^{n} z_i = \frac{2k-n+1}{\sqrt{n+1}}$ par la norme du vecteur z qui est $\|z\| = 1$, notre formulation devient :

$$\begin{aligned} & \max v^t Av; \\ & s.t \|v\| = 1 \\ & v \in \{\frac{-1}{\sqrt{n+1}}, \frac{1}{\sqrt{n+1}}\}^{(n+1)} \end{aligned} \tag{4.23}$$

Le traitement de la formule (4.23) necessite forcément, la réévaluation de la valeur de c dans l'expression (4.13), tant dis que la valeur de B reste la même. Pour avoir une contrainte satisfaite, nous posons :

$$B = \begin{bmatrix} 11.. & .1 \\ 10.. & .0 \end{bmatrix}$$
$$c = \frac{1}{\sqrt{n+1}}(2k-n+1,1)$$

En effet, après ces adaptation et changement de variables, notre formulation est cohérente avec la solution proposée. Une fois que vecteur résultant v est renvoyé, on peut appliquer différents schémas d'arrondissement à $v_{2:n}$ et récupérer la solution relâchée de z ou w, et l'ensemble se caractéristique pertinente peut être sélectionné en fonction de cela. Pour cela, nous conservons simplement les k première caractéristiques ayant les valeurs les plus élevées dans $v_{2:n}$ pour construire l'ensemble optimal. Les résultats des tests sont présentés dans le chapitre suivant.

4.8 Conclusion

Dans ce chapitre nous avons présenté en détaille une nouvelle approche $Eigen\_FS$ de sélection de caractéristiques. Cette approche est basée sur le calcul de valeurs propres de la matrice de performance en tenant compte de la contrainte de cardinalité. $Eigen\_FS$ est une approche hybride qui est basée à la fois sur le classement et sur l'optimisation. L'apport de $Eigen\_FS$ par rapport aux approches de classement est qu'elle tient compte de la relation implicite et explicite entre les caractéristiques. L'apport de $Eigen\_FS$ par rapport aux approches d'optimisation est que le point de départ de la recherche n'est pas aléatoire. Il est important aussi de rappeler que $Eigen\_FS$ prend en compte à la fois la pertinence des caractéristiques et la redondance.

Notre solution proposée comprend plusieurs étapes. Nous avons formulé le problème de sélection de caractéristiques comme un problème d'optimisation générale, nous avons ensuite fait une approximation de notre problème au problème de coupe maximale qui a été relaxé pour s'approcher du problème de calcul de valeurs propre avec contraintes linéaires. Ceci ensuite a été résolu efficacement avec la méthode de puissance projetée. Entre la dernière formulation du problème de calcul de valeurs propre maximales et l'algorithme de la solution, nous sommes passées par quelques adaptation.

Dans le chapitre suivant nous allons présenter le protocole expérimental suivi pour valider notre approche et les résultats obtenus. Les performances de $Eigen\_FS$ ont été validées en utilisant plusieurs métriques d'évaluation des systèmes de classification. De plus notre approche $Eigen\_FS$ a été comparée avec d'autres approches de sélection en utilisant les tests statistiques. Dans un chapitre ultérieur nous présentons le processus de la sélection des caractéristiques visuelle pour la classification des images endoscopiques.

Chapitre 5

Protocole expérimental et résultats expérimentaux

5.1 Introduction

Dans ce chapitre, nous présentons les résultats de l'évaluation de l'approche que nous avons proposée $Eigen\_FS$ en effectuant diverses expériences. En outre, $Eigen\_FS$ est comparée aux méthodes de sélection de l'état de l'art, à savoir la méthode *mRMR*, $InfFS\_s$ et $SSFS$. De plus, nous avons effectué la classification sur les données brutes (avant de faire la sélection), pour voir l'impact de la sélection des caractéristiques sur la classification, cette démarche est référée par $Baseline$.

Afin de valider l'efficacité de $Eigen\_FS$, des expériences approfondies ont été réalisées sur vingt ensembles de données du monde réel provenant du référentiel *UCI Machine Learning* [90]. Nous allons commencer par décrire les ensembles de données utilisés dans notre étude, puis nous donnons un aperçu de la méthodologie adoptée pour mener les expériences et les paramètres d'évaluation. Ensuite, nous présentons nos observations et les résultats des expériences.

5.2 Déscription des ensembles de données

Dans notre étude, vingt ensemble de données issus du référentiel UCI ont été utilisés pour valider l'approche $Eigen\_FS$. Ces jeux de données proviennent de différents domaines, incorporant plusieurs caractéristiques, avec de nombreuses instances et diverses classes. Un résumé détaillé des propriétés des jeux de données est présenté dans **le tableau 5.1** [90].

Comme mentionné dans **le tableau 5.1**, chaque ensemble de données diffère de l'autre non seulement en termes du nombre d'instances, nombre de caractéristiques et nombre de classe mais aussi en termes du domaine d'application.

Dans ce qui suit, nous présentons le protocole expérimental et les résultats.

TABLE 5.1 – Propriétés des ensembles de données utilisés dans les expériences

Dataset	Abréviation	Instances	Caractéristiques	classes
Arrythmia	Arr	452	279	16
Audiology	Aud	226	69	24
Australian Credit Approval	ACA	690	14	2
Balance	Bal	526	4	3
Balloons small	BS	20	4	3
Breast cancer	BC	286	9	2
Breast cancer Wisconsin	BCW	699	10	3
Car evaluation	CE	1728	6	4
Congressional voting records	CVR	435	16	2
Dermatology	Der	366	34	16
Glass Identification	GI	214	10	6
Hepaitis	Hep	155	19	2
Image Segmentaion	IS	2310	19	7
Iris	Ir	150	4	3
Letter recongnition	LR	20000	16	26
Multi-feature fourier	M2F	2000	76	10
Multi-fea karhunen-love	MFKL	2000	64	10
Musk 1	Mu1	476	166	2
Musk 2	Mu2	6598	166	2
Zoo	Zoo	101	16	7

5.3 Stratégie d'expérimentation

Notre méthode de sélection des caractéristiques $Eigen\_FS$ ne concorde pas avec des classifieurs spécifiques. Par conséquent, nous nous attendons à ce que les caractéristiques sélectionnées par ce schéma aient de bonnes performances sur différents types de classifieurs. Afin de tester cela, nous avons considéré deux des algorithmes les plus influents souvent utilisés dans la communauté de la fouille de données ([47]), à savoir machine à vecteurs de support SVM et arbre de décision DT. Comme décrit dans le chapitre 2 les $SVMs$ sont considérées comme les méthodes les plus robustes et les plus précises et les DTs fournissent une méthode efficace pour la prise de décision ([47]).
Nous avons mené et comparé une série d'expériences pour évaluer la classification basée sur DT et SVM. La méthodologie adoptée dans nos expériences est décrite ci-dessous :

— Appliquer la classification avec DT et SVM sur les ensembles de données, avant d'effectuer la sélection des caractéristiques *Baseline* et se référer aux résultats obtenus pour comparer l'efficacité de notre approche.

— Effectuez une série de tests de notre approche $Eigen\_FS$, en variant à chaque fois le pourcentage du nombre de caractéristiques sélectionnées (λ). Sur chaque sous-ensemble de caractéristiques obtenu, nous appliquons des classifications DT et SVM et évaluons leurs efficacités (Classification basée sur $Eigen\_FS$).

— Effectuez une série de tests de la méthode *mRMR* [16],en variant à chaque fois le pourcentage du nombre de caractéristiques sélectionnées (λ). Sur chaque sous-ensemble de caractéristiques obtenu, nous appliquons *DT* et *SVM* et évaluons leurs efficacités (Classification basée sur *mRMR*).

— Effectuez une série de tests de la méthode $InfFS_s$ [86], en variant à chaque fois le pourcentage du nombre de caractéristiques sélectionnées (λ). Sur chaque sous-ensemble de caractéristiques obtenu, nous appliquons *DT* et *SVM* et évaluons leurs efficacités (Classification basée sur $InfFS_s$).

— Effectuez une série de tests de la méthode *SSFS* [88], en variant à chaque fois le pourcentage du nombre de caractéristiques sélectionnées (λ). Sur chaque sous-ensemble de caractéristiques obtenu, nous appliquons *DT* et *SVM* et évaluons leurs efficacités (Classification basée sur *SSFS*).

5.4 Environnement expérimental

L'approche proposée et toutes les approches concurrentes sont mises en œuvre en utilisant Matlab $R2016b$. Toutes les expériences sont exécutées sur un système équipé d'un processeur Intel Core $i5-93000H$, $2,40GHzx8$, $16Go$ de RAM avec partition d'échange et d'un GPU GeForce $GTX1650/PCIe/SSE2$.

5.5 Détails techniques des éxpériences

En raison du petit nombre limité d'échantillons, nous avons opté pour la stratégie de validation croisée qui permet d'utiliser l'ensemble de données pour l'apprentissage et la validation simultanément. Dans nos exécutions, nous avons divisé l'ensemble de données en k parties (folds) et pour chaque itération, nous avons sélectionné une partie de k, en l'utilisant comme partie de test "test". Le reste (l'union des $k-1$ autres parties) a été utilisé pour l'apprentissage "train". Toutes les expériences ont été réalisées par la méthode de validation croisée 5×2. Notre modèle a été formé à l'aide de la partie "train" et testé à l'aide de la partie "test". Par conséquent, nous avons obtenu cinq ensembles formés et 5 estimations d'exactitude "Accuracy" de chaque test. Enfin, nous avons considéré uniquement la moyenne de ces cinq estimations. Les valeurs moyennes de ces cinq estimations de précision ont été considérées comme les résultats finaux de l'expérience.

Dans cette étude, nous avons évalué les modèles prédictifs, utilisant les classifieurs mentionnés ci-dessus, sur la base de différents critères d'évaluation, à savoir : l'exactitude, la précision et le score F1. Les mesures utilisées pour évaluer notre modèle sont faites pour des problèmes de classification binaire. Cependant, elles peuvent être étendues aux problèmes de classification multiple en considérant une des classes comme positive et les autres comme négatives. La

moyenne de ces mesures pour les classes individuelles devient alors les valeurs finales du modèle entier. Nous avons choisi d'utiliser l'exactitude pour mettre en évidence les TP et TN, donc pour déterminer l'efficacité du système. En complément de l'exactitude, le score F1 est utilisé pour mettre en évidence les FN et FP. Enfin, la précision est utilisée pour déterminer quand les coûts de la FP sont élevés.

5.6 Resultats et discussion

Dans cette section, les résultats obtenus par chaque classifieur sont présentés. Nous commençons par les résultats de la classification DT et ensuite, nous présentons les résultats obtenus par le classifieur SVM.

5.6.1 Résultats de la classification DT

Cette expérience avait pour but de comparer les résultat de la classification DT de base, c'est à dire avec les données brutes, avant d'appliquer la sélection des caractéristiques (*baseline*) avec DT basé sur $Eigen\_FS$ et avec DT basé sur *mRMR* , $InFS_s$ et $SSFS$. L'exactitude (accuracy) de la classification est indiquée dans le **Tableau 5.2**. L'accuracy maximale et l'accuracy moyenne des résultats obtenus en variant le nombres de caractéristiques sélectionnées sont indiquées.

TABLE 5.2 – Résumé des résultats de l'accuracy moyenne obtenus par la classification *DT*

Dataset	mRmr			*InfFS_S*			SSFS			*Eigen_FS*			*DT_Baseline*
	Moyenne	Max	λ	Moyenne	Max	Lambda	Moyenne	Max	λ	Moyenne	Max	λ	
CE	82.51	86.82	0.7	68.47	70.02	0.2	79.15	80.93	0.5	82.51	86.82	0.7	**94.24**
BC	67.8	69.44	0.5	69.53	70.84	0.3	66.23	67.27	0.3	70.85	**73.21**	0.4	65.10
M2F	73.03	74.17	0.2	72.44	73.37	0.2	70.01	71.53	0.3	73.13	**74.91**	0.2	71.35
Ir	94.34	95.6	0.8	77	94.13	0.5	93.81	94.93	0.5	95.86	**96**	0.5	95.6
Arr	63.01	63.67	0.3	60.26	62.88	0.4	59.70	63.14	0.7	63.53	**64.47**	0.2	62.47
Aud	72.33	72.42	0.3	61.96	67.03	0.3	71.94	72.3	0.7	72.33	**72.42**	0.3	71.85
MFKL	78.39	**79.57**	0.4	78.28	79.08	0.4	78.56	97.1	0.4	78.56	97.3	0.3	77.83
BCW	94.53	94.82	0.5	93.74	94.76	0.3	95.93	**96.42**	0.6	94.19	94.47	0.6	94.04
LR	81.73	83.10	0.7	66.84	81.88	0.7	78.87	**83.11**	0.7	80.38	83.10	0.7	82.31
BS	57.14	57.14	0.3	57.14	57.14	0.3	57.14	57.14	0.3	57.14	57.14	0.3	57.14
ACA	80.01	80.61	0.5	82.16	82.63	0.5	72.03	80.04	0.7	84.66	**87.06**	0.2	80.78
CVR	94.98	95.03	0.7	94.01	94.13	0.7	92.83	94.09	0.6	95.10	**96.9**	0.7	94.03
Der	87.82	93.28	0.7	85.90	88.97	0.7	89.31	91.48	0.7	87.16	**94.36**	0.7	93.2
Hep	81.34	82.12	0.4	77.17	79.42	0.3	77.15	77.31	0.3	83.36	**84.5**	0.6	76.92
GI	51.9	56.47	0.6	53.02	56.76	0.7	51.22	55.33	0.7	53.17	**58.62**	0.6	56.64
Mus1	71.04	71.38	0.5	71.43	72.52	0.7	68.55	70.82	0.7	71.43	**73.01**	0.7	70.91
Mus2	92.82	93.42	0.7	92.09	92.49	0.6	90.95	92.89	0.5	93.64	**94.84**	0.5	93.03
IS	91.62	93.64	0.6	93.03	**96.65**	0.6	91.35	92.66	0.7	90.89	92.77	0.7	92.75
Zo	81.99	82.94	0.6	82	82.65	0.6	81.70	82.94	0.5	82.79	**83.94**	0.5	83.24
Bal	64.22	65.66	0.5	61.81	62.85	0.5	63.5	65.76	0.5	67.56	70.25	0.5	**79.75**

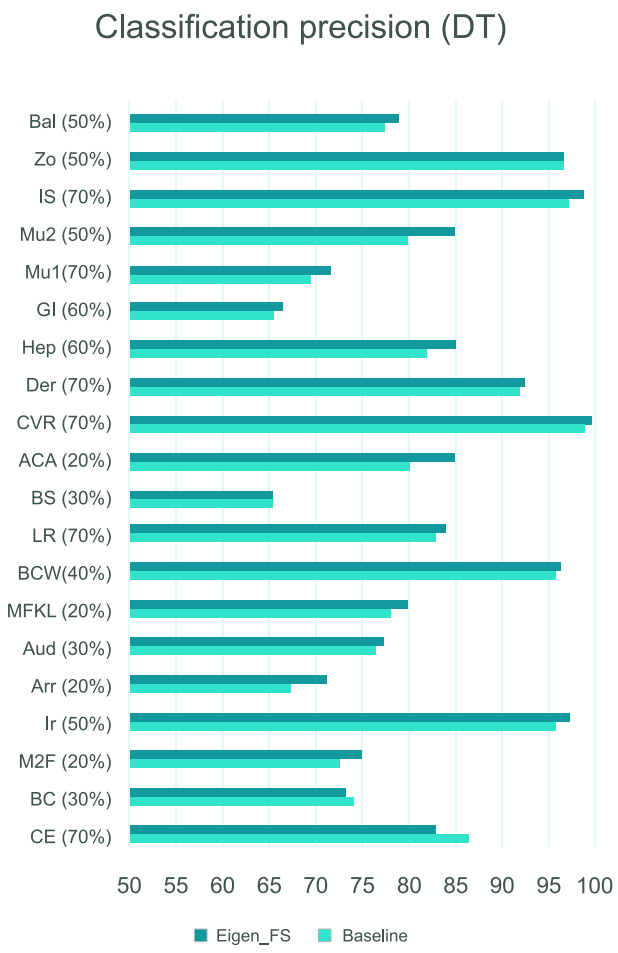

FIGURE 5.1 – Résumé des résultats de la précision moyenne obtenus par la classification *DT*.

Nous observons à partir de **Table 5.2** que les résultats de la classification *DT* après la sélection basée sur *Eigen_FS* ont été améliorés dans tous les cas, comparés aux résultats de la classification après la sélection basé sur *mRMR* , $InFS_s$ et *SSFS*. Les résultats ont été amélioré aussi dans la plupart des cas comparés aux résultats de la classification "baseline". Notamment, pour l'ensemble Breast Cancer, les ACCs obtenus avant et après la sélection sont **65,10%** and **73,21%** respectivement contre **69.44%**, **69.53%**, et **66.23%** obtenus par *mRMR*, *InfFS_FS*, et *SSFS* respectivement.

Figure 5.1 présente les résultats de précision de la classification *DT* en considérant les caractéristiques sélectionnées par *Eigen_FS* de tous les jeux de données. La précision maximale obtenue par *DT* basée sur *Eigen_FS* et la précision atteinte par baseline *DT* sont indiquées. Le pourcentage du nombre de caractéristiques sélectionnées est indiqué après le nom de l'ensemble de données. Nous observons que le *DT* basé sur *Eigen_FS* surpasse le baseline *DT* en termes de précision, dans tous les cas, sauf dans le cas des ensembles de données BC et CE. Les valeurs

TABLE 5.3 – Résultats du score F1 et du taux d'erreur obtenus par la classification *DT*.

Dataset	F1_Score	Erreur	λ
CE	0.72	0.12	0.7
BC	0.85	0.12	0.4
M2F	0.75	0.15	0.3
Ir	0.98	0.01	0.5
Arr	0.79	0.28	0.2
Aud	0.82	0.19	0.6
MFKL	0.83	0.19	0.3
BCW	0.96	0.04	0.6
LR	0.84	0.04	0.15
BS	0.62	0.29	0.3
ACA	0.86	0.12	0.2
CVR	0.96	0.04	0.7
Der	0.93	0.06	0.7
Hep	0.61	0.19	0.6
GI	0.78	0.29	0.6
Mu1	0.77	0.15	0.7
Mu2	0.81	0.07	0.5
IS	0.94	0.05	0.7
Zo	0.88	0.14	0.5
Bal	0.72	0.21	0.5

les plus remarquables sont obtenues sur les ensembles Arythmia et Australian Credit Approval, de sorte que les précisions obtenues par $Eigen\_FS$ sont respectivement de **77 %** et **85 %**, contre **68%** et **80 %** respectivement obtenus par la classification baseline.

En plus de l'accuracy et de la précision, le tableau 5.3 montre les meilleures valeurs du score F1 et de l'erreur [1] obtenus par DT basé sur $Eigen\_FS$.

5.6.2 Résultats de la classification *SVM*

Les résultats en termes d'accuracy de la classification *SVM* sont résumés dans **Table 5.4**, **Table 5.5** et **Table 5.6**. En considérant 3 différentes fonctions de noyau, à savoir "linear", "gaussian" et "polynomial". L'accuracy maximale et l'accuracy moyenne des différents sous-ensembles (de différente taille) de caractéristiques sélectionnées sont présentées.

Nous pouvons observer à partir des résultats présentés dans le **Tableau 5.4** obtenus par la classification *SVM* avec une fonction "linear", une amélioration notable de la classification après une sélection basé sur $Eigen\_FS$. Notre approche améliore l'accuracy de la classification par rapport aux sélections basés sur *mRMR* dans tous les cas. Elle est également plus performante que la classification *SVM* de base dans la plupart des cas, en particulier pour l'ensemble

de données Audiology ; ainsi, la meilleure accuracy obtenue avant et après la sélection des caractéristiques est de **70,97%** et de **75,86%** respectivement, contre **74,78%**, **67,03%** et **73,69%** obtenus par *mRMR*, $InfFS\_FS$ et $SSFS$ respectivement.

TABLE 5.4 – Résumé des résultats de l'accuracy moyenne obtenus par la classication *SVM* en utilisant le noyau 'Linear'

Dataset	mRmr			*InfFS_S*			SSFS			*Eigen_FS*			*SVM_Baseline*
	Moyenne	Max	λ	Moyenne	Max	λ	Moyenne	Max	λ	Moyenne	Max	λ	
CE	70.08	70.17	0.7	68.99	70.02	0.2	70.13	70.57	0.7	70.08	70.17	0.7	**70.45**
BC	70.32	70.83	0.8	70.09	70.91	0.7	69.72	70.07	0.5	70.32	70.69	0.5	**71.25**
M2F	81.27	84.66	0.3	83.38	84.51	0.4	80.01	81.62	0.3	81.73	**84.82**	0.2	81.38
Ir	95.57	96.53	0.8	78.86	95.20	0.5	95.05	94.53	0.3	96.13	97.2	0.8	**97.33**
Arr	64.17	65.70	0.2	65.15	**66.64**	0.6	63.80	65.53	0.7	63.75	65.15	0.1	63.89
Aud	71.40	74.78	0.3	61.96	67.03	0.3	72.73	73.69	0.3	71.78	**75.86**	0.3	70.97
MFKL	95.73	**95.95**	0.8	94.27	95.51	0.7	95.15	95.6	0.6	95.70	96.01	0.8	**96.15**
BCW	95.89	96.28	0.6	94.77	96.57	0.7	95.83	**96.57**	0.7	96.12	96.45	0.7	96.22
LR	75.82	81.72	0.8	54.96	78.67	0.7	69.21	80.6	0.7	74.83	81.72	0.8	**85.41**
BS	76.85	**84.29**	0.5	73.42	77.14	0.5	75.14	81.43	0.5	78.63	83.43	0.5	82.86
ACA	85.6	85.92	0.5	85.66	85.65	0.4	77.09	85.96	0.6	85.52	**86.52**	0.2	85.3
CVR	94.25	94.90	0.6	94.36	94.83	0.3	94.44	94.69	0.4	95.24	**96.21**	0.5	94.97
Der	91.65	95.82	0.7	& 92.4	95.00	0.5	92.45	94.98	0.7	89.42	**95.08**	0.7	92.21
Hep	82.38	82.5	0.5	82.06	82.88	0.4	80.61	81.73	0.3	82.13	**83.27**	0.4	79.23
GI	46.27	52.05	0.7	50.39	52.69	0.3	53.20	59.16	0.6	50.97	**62.59**	0.7	59.03
Mus1	75.43	77.8	0.7	76.41	78.43	0.5	76.65	**78.99**	0.6	75.40	78.11	0.7	72.69
Mus2	92.59	93.45	0.6	92.73	93.88	0.7	92.93	94.02	0.7	93.58	**96.22**	0.7	94.86
IS	88.30	94.25	0.7	93.34	**93.94**	0.5	**90.80**	94.53	0.7	89.78	**96.99**	0.5	94.42
Zo	89.17	92.06	0.7	89.52	90.29	0.7	88.64	91.18	0.7	88.64	91.47	0.7	**92.35**
Bal	67.12	76.61	0.5	64.62	66.39	0.5	65.30	67.82	0.5	65.42	68.01	0.5	**88.59**

TABLE 5.5 – Résumé des résultats de l'accuracy moyenne obtenus par la classication *SVM* en utilisant le noyau 'Gaussian'

Dataset	mRmr			*InfFS_S*			SSFS			*Eigen_FS*			*SVM_Baseline*
	Moyenne	Max	Lambda	Moyenne	Max	Lambda	Moyenne	Max	Lambda	Moyenne	Max	Lambda	
CE	82.37	86.56	0.7	68.99	70.02	0.7	79.06	86.56	0.7	82.37	86.56	0.7	**89.01**
BC	70.43	71.53	0.5	71.03	72.17	0.5	69.39	70	0.5	72.78	**74.35**	0.4	69.02
M2F	31.83	60.27	0.2	29.3	60.83	0.2	19.29	31.54	0.3	31.87	**60.62**	0.2	11.31
Ir	94.26	95.33	0.8	82.20	92.94	0.5	94.67	94.67	0.5	96.15	97.13	0.5	94.53
Arr	55.37	**61.15**	0.02	54.20	54.20	0.2	54.2	54.02	0.4	55.42	60.95	0.02	& 54.20
Aud	40.19	58.91	0.1	61.96	**67.03**	0.3	37.37	40.8	0.3	40.28	59.78	0.09	& 33.36
MFKL	16.13	**63.29 &**	0.1	30.40	72.94	0.2	23.96	45.71	0.3	36.44	88.08	0.1	10.78
BCW	95.90	**96.39**	0.5	94.95	96.08	0.5	95.94	**96.14**	0.5	95.93	96.25	0.5	94.50
LR	90.92	93.59	0.7	74.33	92.35	0.6	87.16	**93.6**	0.7	89.85	93.59	0.7	87.80
BS	79.82	**88.57**	0.5	74.28	78.57	0.5	75.14	81.43	0.4	81.71	**88.57**	0.5	82.86
ACA	80.95	84.70	0.4	83.85	84.74	0.4	74.53	81.78	0.5	82.84	**86.03**	0.2	63.26
CVR	92.63	94.62	0.4	90.86	94.69	0.3	92.81	**96.41**	0.3	91.82	93.93	0.3	79.66
Der	52.77	72.82	0.3	40.75	53.85	0.3	43.62	61.56	0.3	55.06	**73.93**	0.3	30.33
Hep	79.00	**82.12**	0.3	78.81	78.85	0.4	78.92	79.23	0.3	78.77	**82.12**	0.3	78.85
GI	57.08	59.71	0.6	55.54	**60.69**	0.5	56.43	59.43	0.7	51.97	59.29	0.7	56.35
Mus1	60.05	67.17	0.3	57.02	60.57	0.3	57.03	**58.24**	0.3	61.43	**70.05**	0.3	56.6
Mus2	88.25	90.24	0.3	87.96	88.87	0.3	87.63	87.65	0.6	88.58	**91.44**	0.2	87.64
IS	91.38	93.34	0.5	93.23	**93.42**	0.5	91.49	90.75	0.6	90.21	92.38	0.4	85.13
Zo	85.82	86.76	0.5	79.80	81.97	0.4	86.58	**88.82**	0.3	87.17	88.24	0.4	**57.35**
Bal	65.06	66.71	0.7	65.15	66.67	0.5	65.26	67.05	0.5	66.95	71.43	0.3	**82.86**

TABLE 5.6 – Résumé des résultats de l'accuracy moyenne obtenus par la classication *SVM* en utilisant le noyau 'Polynomial'

Dataset	mRmr			*InfFS_S*			SSFS			*Eigen_FS*			*SVM_Baseline*
	Moyenne	Max	λ	Moyenne	Max	λ	Moyenne	Max	λ	Moyenne	Max	λ	
CE	82.46 &	86.99	0.7	69.21	70.02	0.2	78.67	84.99	0.7	82.46	86.99	0.7	**95.67**
BC	68.30 &	70.27	0.4	70	71.82	0.3	66.50	69.09	0.3	70.62	**73.42**	0.5	61.95
M2F	79.30 &	83.32	0.3	81.26	82.38	0.2	75.22	78.71	0.3	79.30	**83.51**	0.2	76.07
Ir	95.28	& 95.46	0.5	80.14	93.66	0.5	94.8	96.18	0.4	96.18	96.26	0.4	92.93
Arr	56.37 &	**61.63**	0.3	54.85	59.73	0.4	53.55	58.5	0.3	54.24	61.68	0.3	21.72
Aud	63.60 &	66.93	0.3	61.96	67.03	0.3	63.85	67.23	0.3	63.55	**68.40**	0.3	56.46
MFKL	94.73 &	**95.4**	0.3	94.28	95.54	0.5	95.30	95.84	0.4	94.98	**95.91**	0.3	93.87
BCW	94.74 &	**95.59**	0.4	93.86	95.48	0.3	94.93	95.57	0.3	94.63	95.02	0.4	93.47
LR	90.59 &	93.76	0.7	72.76	92.81	0.7	85.93	9307	0.7	89.34	93.76	0.7	**93.98**
BS	79.42	**88.57**	0.5	73.21	75.71	0.5	75.14	71.43	0.5	81.71	**88.57**	0.5	81.43
ACA	79.41	84.43	0.4	80.75	83.93	0.3	71.84	76.74	0.5	82.27	**85.57**	0.2	77.26
CVR	93.95	94.41	0.6	93.69	94.1	0.3	95.72	**96.48**	0.3	94.52	94.9	0.5	95.66
Der	86.88	91.89	0.7	91.19	92.13	0.7	89.64	92.38	0.7	87.49	**92.31**	0.7	90.64
Hep	78.17	**81.05**	0.3	79.52	81.35	0.7	77.88	80.96	0.7	80.11	**82.5**	0.2	81.92
GI	53.79	60.97	0.7	57.88	59.13	0.5	54.25	57.33	0.7	54.68	**61.94**	0.7	55.06
Mus1	78.09	82.14	0.7	80.50	81.96	0.7	78.69	**80.31**	0.7	79.94	**82.89**	0.7	**83.85**
Mus2	95.54	96.08	0.5	95.42	95.79	0.7	94.56	96.61	0.7	95.56	96.95	0.5	97.28
IS	93.00	95.16	0.6	95.48	**95.79**	0.7	94.13	95.26	0.6	92.9	94.73	0.5	92.7
Zo	89.81	90.88	0.5	90.15	**95.71**	0.4	89.82	90.59	0.6	90.39	91.08	0.4	88.82
Bal	63.41	65.67	0.5	64.10	64.06	0.5	64.20	66.52	0.5	64.32	66.72	0.5	**72.33**

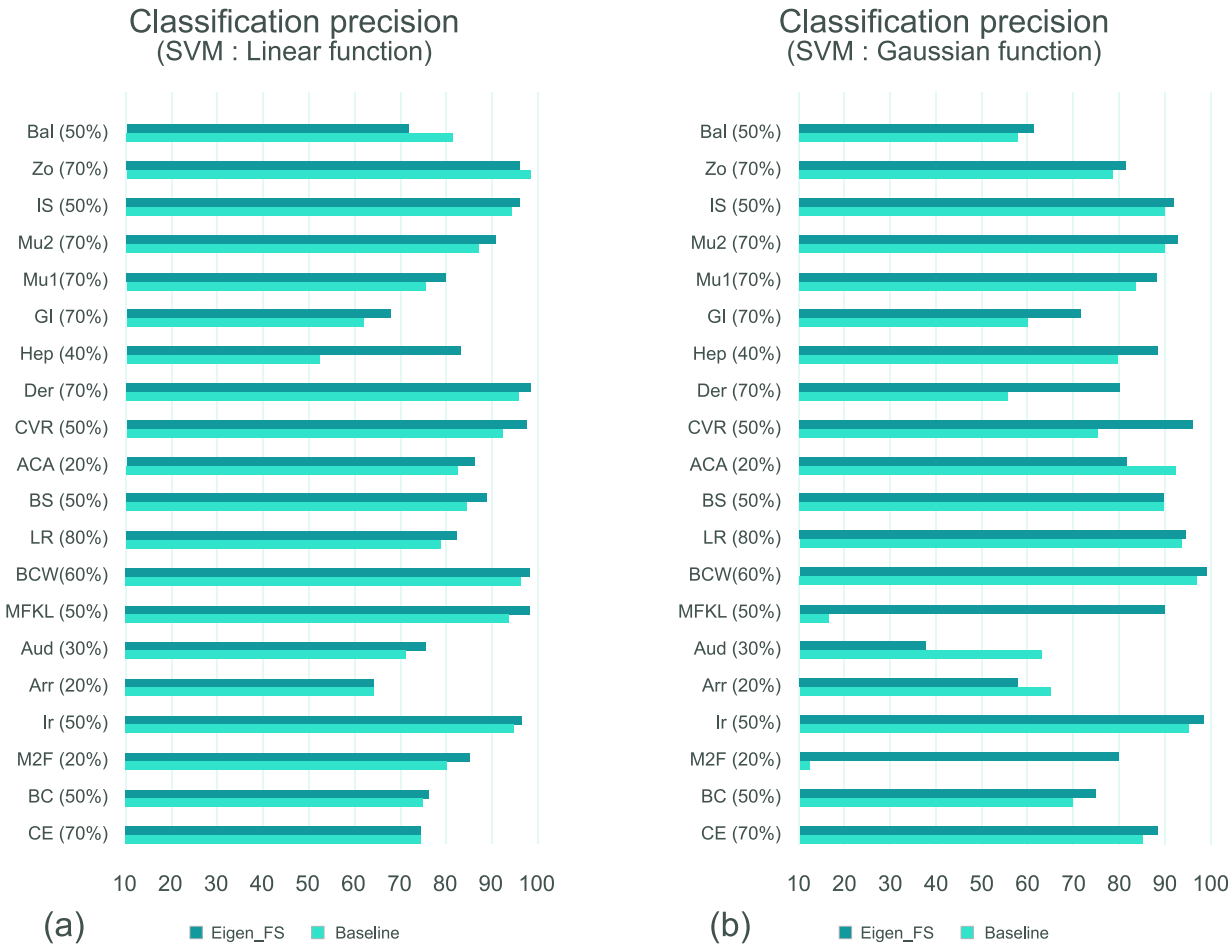

FIGURE 5.2 – Précision de la classification *SVM*

Le tableau 5.5 présente les résultats obtenus par la classification *SVM* avec la fonction 'gaussian'. D'après ce tableau, nous observons que *SVM* basé sur la sélection *Eigen_FS* a de meilleures performances que *SVM* basé sur *mRMR* dans tous les cas. De plus, la classification utilisant *SVM* basé sur *Eigen_FS* est plus performante que la classification *SVM* de base dans la plupart des cas, en particulier pour l'ensemble de données Multi Feature Fourier ; ainsi, la meilleure accuracy obtenue avant et après la sélection des caractéristiques est de **11,31%** et **60,62 %** respectivement, contre **60,27%, 60,83%** et **31,54%** obtenus par *mRMR*, $InfFS\_FS$ et *SSFS* respectivement.

Tableau 5.6 présente les résultats de la classification *SVM* obtenus avec la fonction 'Polynomial'. D'après ce tableau, on remarque que *SVM* basé sur *Eigen_FS* est plus performant que *SVM* basé sur *mRMR* dans tous les cas. Les résultats révèlent également que la classification basée sur notre approche est meilleure que la classification basée sur le *SVM* de base dans la

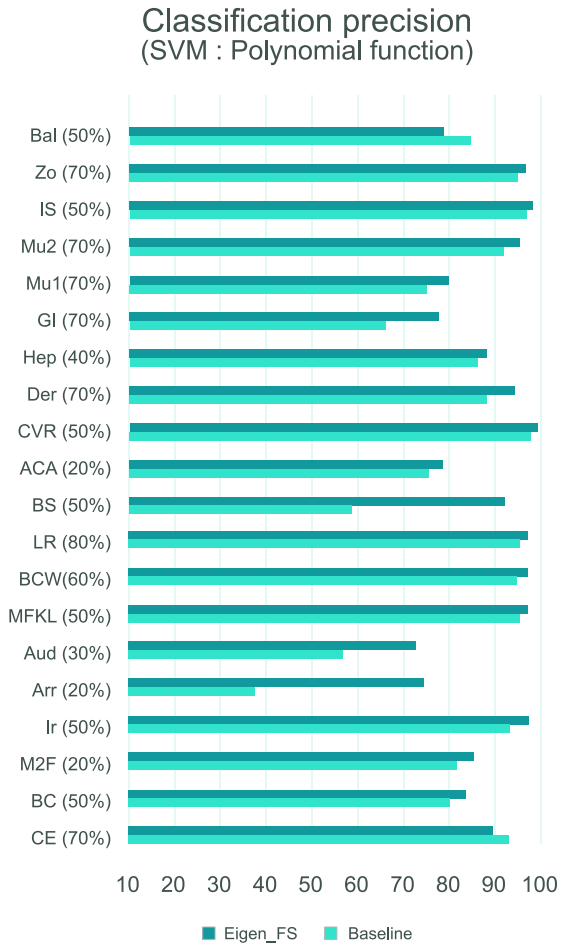

FIGURE 5.3 – Précision de la classification *SVM*

plupart des cas, en particulier pour l'ensemble de données Breast Cancer. Les meilleures accuracies obtenues avant et après la sélection des caractéristiques sont respectivement de **61,95%** et **73,42%**, contre **70,27%**, **71,82%**, et **69,09%** obtenus par *mRMR*, $InfFS\_FS$, et *SSFS* respectivement.

Figure 5.2 et **Figure 5.3** illustrent les résultats de la précision de classification *SVM* en considérant les caractéristiques sélectionnées par $Eigen\_FS$ de tous les ensemble de données. La précision maximale atteinte par *SVM* basé sur $Eigen\_FS$ et la précision obtenue par le *SVM* de base sont rapportées. Le pourcentage du nombre de caractéristiques sélectionnées est indiqué après le nom de l'ensemble de données.

Il est intéressant de mentionner que le *SVM* basé sur $Eigen\_FS$ surpasse le classifieur *SVM* de base en termes de précision. Les valeurs les plus remarquables sont atteintes sur les ensembles de données Audiology, Arythmia, Multi-Features-Fourier et Multi-Features-Kernel-Love. En outre, le classifieur de base *SVM* obtient de meilleures performances avec le noyau 'polynomial',

TABLE 5.7 – Résultats du score F1 et du taux d'erreur obtenus par la classification *SVM*

Dataset	F1_Score	Erreur	λ	Kernel
CE	69.00	0.12	0.7	Polynomial
BC	85.00	0.23	0.4	Gaussian
M2F	75.00	0.22	0.2	Linear
Ir	98.00	0.02	0.5	Linear
Arr	82.00	0.29	0.1	Linear
Aud	89.00	0.18	0.3	Linear
MFKL	95.00	0.04	0.5	Linear
BCW	97.00	0.03	0.7	Linear
LR	94.00	0.05	0.7	Polynomial
BS	80.01	0.12	0.5	Gaussian
ACA	87.02	0.12	0.2	Linear
CVR	94.12	0.07	0.5	Linear
Der	97.02	0.03	0.7	Linear
Hep	63.89	0.13	0.4	Linear
GI	53.06	0.39	0.7	Linear
Mu1	85.82	0.13	0.7	Polynomial
Mu2	90.96	0.03	0.5	Polynomial
IS	92.94	0.07	0.5	Linear
Zo	90.03	0.07	0.4	Polynomial
Bal	72.96	0.21	0.3	Gaussian

dans la plupart des cas. Les précisions atteintes par *SVM* en utilisant la fonction 'lineair' pour l'ensemble de données Hepaitis sont de **92,3%** et **52,29%** par $Eigen\_FS$ et l'approche de base respectivement. Les précisions atteintes par *SVM* utilisant la fonction 'gaussian' pour l'ensemble de données Multi-fea karhunen-love sont respectivement de **90%** et **17,36%** par $Eigen\_FS$ et l'approche de base. Les précisions atteintes par *SVM* utilisant la fonction polynomiale pour l'ensemble de données Arrythmia sont **74,89%** et **37,91%** par $Eigen\_FS$ et l'approche de base respectivement. En plus de l'exactitude et de la précision, **le tableau 5.7** montre les meilleures valeurs du score F1 et de l'erreur ([1]) obtenus par *SVM* basé sur $Eigen\_FS$.

5.7 Tests statistiques

En plus des précédents tests, nous avons comparé les performances de notre approche et des approches compétitives en utilisant les rangs moyens sur les 20 ensembles de données. Suivant les recommandations de Demšar ([52], [101]) , nous avons d'abord effectué un test de Friedman pour comparer statistiquement les performances de ces approches. De plus, puisque nous sommes seulement intéressés par comparer l'approche Baseline avec toutes les autres techniques de sélection de caractéristiques, nous avons enchainé par un test de Bonferroni-Dunn en considérant *Baseline* comme l'approche de contrôle.

TABLE 5.8 – Rang moyen des 5 techniques de sélection comparées

mRMR	InfFS_s	SSFS	*Eigen_FS*	Baseline
2.5250	3.6500	3.7250	1.6000	3.5000

Figure 5.4 montre les résultats du test de Bonferroni-Dunn à un niveau de signification de $0,1\%$ avec la valeur critique $q_{0,001} = 3,66$ et la différence critique $CD = 1,83$. Sur l'axe horizontal, nous représentons les rangs moyens de chaque méthode (donnés dans le **tableau 5.8**) et nous marquons à l'aide d'une ligne épaisse l'intervalle d'un *CD* à gauche et à droite du rang moyen de Baseline. Toute technique dont le rang se situe en dehors de cette zone est significativement différente de la méthode de contrôle.

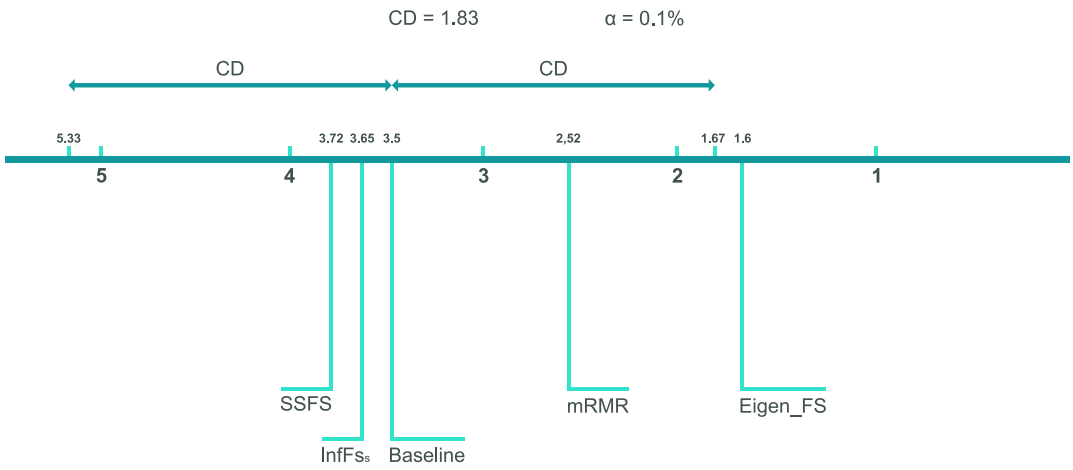

FIGURE 5.4 – Comparaison de l'approche Baseline avec les autres modèles de sélection avec le test de Dumn de Bonferroni

L'analyse des résultats du test de Bonferroni-Dunn illustré par la **figure 5.4** indique que *Eigen_FS* a le rang le plus bas et se situe en dehors de l'intervalle marqué. Par conséquent, nous pouvons conclure que *Eigen_FS* améliore significativement la technique de base, alors que les approches comparées ne parviennent pas à atteindre les mêmes résultats à $0,1\%$ de différence de signification.

Pour une analyse plus approfondie de ces résultats, nous avons comparé ces scores par paires sur la base du test de Wilcoxon dans le **tableau 5.9**. La première ligne de chaque entrée indique le nombre de victoires/égalités/pertes de la technique de la colonne par rapport à la technique de la ligne, tandis que la deuxième ligne indique les valeurs p du test de Wilcoxon. Si l'entrée est en gras, cela signifie que le nombre de victoires/pertes sur 20 est statistiquement significatif en utilisant le test de Wilcoxon.

Les résultats présentés dans le **tableau 5.9** indiquent que la méthodologie proposée est plus performante que les autres alternatives dans la plupart des cas. Plus important encore, notre

TABLE 5.9 – Résumé du test de signe de Wilcoxon

		mRMR	InfFS_s	SSFS	*Eigen_FS*	Baseline
mRMR	W/T/L		5/1/14	2/3/15	**13/4/3**	5/3/12
	p_{value}		0.0731	0.0013	**0.0032**	0.0930
InfFS_s	W/T/L			11/1/8	**17/1/2**	9/1/10
	p_{value}			0.8083	**0.0011**	0.9553
SSFS	W/T/L				**17/2/1**	11/1/8
	p_{value}				**4.4934e-04**	0.6012
Eigen_FS	W/T/L					2/1/17
	p_{value}					0.0117

approche *Eigen_FS* améliore significativement les performances de la *Baseline* avec $p\_value = 0.0117$.

5.8 Conclusion

Dans ce chapitre nous avons présenté en détail la stratégie utilisée pour valider l'approche que nous avons proposée *Eigen_FS*. Nous avons commencé par décrire l'environnement d'exécution. Ensuite nous avons présenté les résultats de test de la classification *DT* basée sur *Baseline*, *Eigen_FS*, *mRMR* , $InFS_s$ et *SSFS*, ceci en termes d'exactitude, précision, score F1 et l'erreur. Nous avons ensuite présenté les résultats de test de la classification *SVM* basée sur *Baseline*, *Eigen_FS*, *mRMR* , $InFS_s$ et *SSFS* en considérant 3 différentes fonction de noyau, ceci en termes d'exactitude, précision, score F1 et l'erreur. Les résultats ont montré que notre proposition *Eigen_FS* améliore significativement les résultats de la classification par rapport à la classification basée sur *Baseline*, *mRMR* , $InFS_s$ et *SSFS* dans la plupart des cas. Pour renforcer nos conclusions et montrer que notre approches est significativement meilleures que les autres, nous avons enchainé par des tests statistiques, ce qui a affirmé l'efficacité de notre approche par rapport aux autres, surtout par rapport à l'approche de base.

Chapitre 6

Sélection des caractéristiques visuelles basée sur $Eigen\_FS$ pour la détection du saignement dans le tractus gastro-intestinal

6.1 Introduction

Lorsque les cellules saines de la paroi du tractus gastro-intestinal (GI) se développent hors du contrôle normal, quelque chose d'anormal se produit, comme un saignement rectal, une inflammation ou une formation de masses [10, 102, 103]. Lorsque les anomalies sont détectées à un stade précoce, la guérison est souvent possible [10, 104]. Par conséquent, pour sauver la vie du patient, nous devons trouver les cellules affectées à un stade précoce. Les techniques endoscopiques conventionnelles, telles que la gastroscopie et la coloscopie ont du mal à montrer l'intestin grêle d'un côté et sont très douloureuses d'un autre. Par conséquent, en raison de leur nature complexe, ces instruments ne sont pas adaptés à l'identification et à l'examen des anomalies gastro-intestinales. En 2000, l'endoscopie par capsule sans fil "Wireless Capsule Endoscopy" (WCE) a été développée pour résoudre le problème des instruments de gastroscopie [10, 103]. WCE est couramment utilisée pour détecter les anomalies dans des conditions indolores et non invasives [10]. Il s'agit d'un dispositif en forme de pilule qui comprend une caméra, une batterie, des sources de lumière et un émetteur radio [10, 105]. Le but de l'utilisation du WCE est de détecter toutes les anomalies du tube digestif. La procédure WCE utilise une minuscule caméra sans fil pour capturer le tube digestif, puis transmet les images à un enregistreur de données [10, 105, 106]. L'intestin humain est examiné, après avoir avalé la capsule. La durée de cette opération est d'environ 8 heures, à raison de 2 images par seconde [10]. Cependant, seulement 5% de ces images peuvent contenir une anomalie chez les patients souffrant d'une maladie gastro-intestinale [10].Dans ce cas, la détection des anomalies est une tâche qui prend beaucoup de temps pour le médecin qui doit analyser visuellement toutes les images. Par conséquent, un système assisté par ordinateur est envisagé pour résoudre ce problème. Depuis 2001, il y a eu une explosion de la médecine littérature concernant le WCE. Les recherches pertinentes sur la

WCE englobent plusieurs sujets majeurs : la segmentation vidéo, l'amélioration de l'image et la détection de maladies. Nous avons également fait le travail pour la détection automatisée des maladies dans les images endoscopiques. Dans notre travail, nous nous intéressons particulièrement à la détection du saignement dans l'appareil digestif, qui est un symptôme commun à de nombreuses maladies gastro-intestinales. Un échantillon d'image normale et d'image de saignement est présenté dans la **figure 6.1 (a)** et **(b)** respectivement. Dans le contexte de la détection automatique du saignement, une première tentative a été faite en fournissant un logiciel avec la capsule endoscopique de deuxième génération, à savoir l'indicateur de sang suspect (SBI) [107]. Ce logiciel offre une précision modérée et ne réduit pas le temps d'interprétation nécessaire. C'est pourquoi les chercheurs sont motivés pour explorer les techniques d'apprentissage automatique pour la détection automatique du saignement dans les vidéos de WCE.

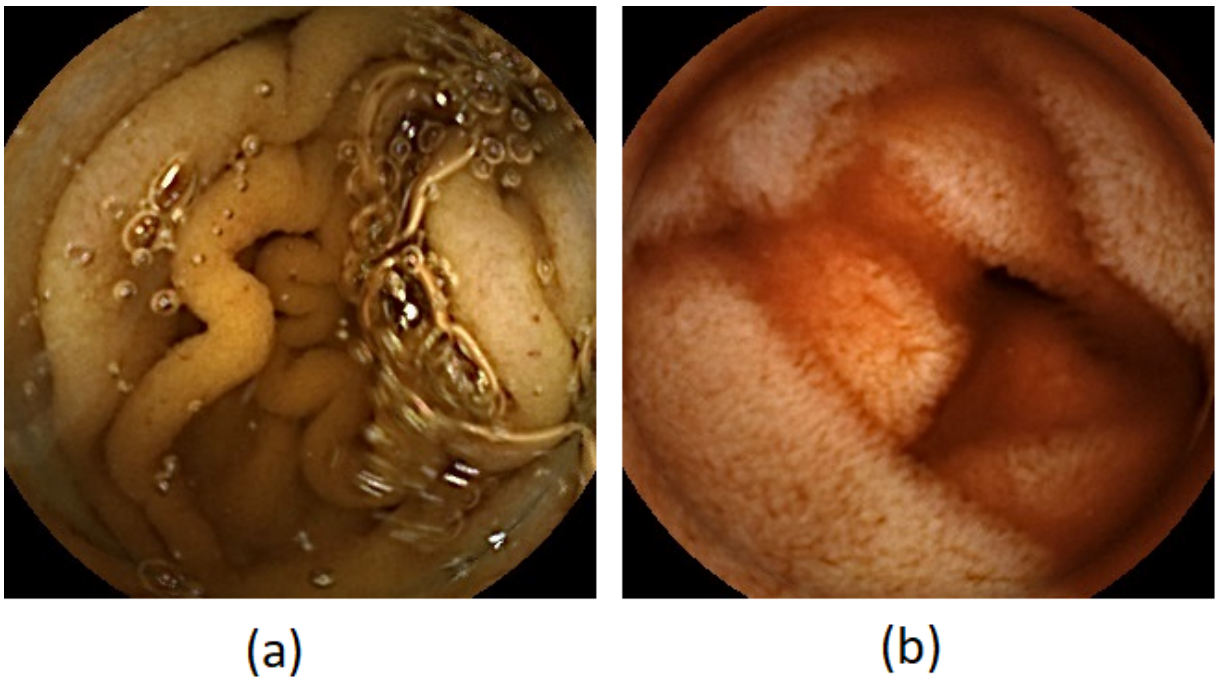

FIGURE 6.1 – (a) image normale, and (b) image avec saignement dans WCE.

Dans ce chapitre, nous proposons un schéma complet de détection d'images endoscopiques contenant du saignement. Ce schéma comprend un prétraitement de l'image, la spécification de la région d'intérêt (ROI), l'extraction de caractéristiques à partir des ROIs, la sélection de caractéristiques (FS) et la classification. La principale contribution est l'exploitation de notre approche précédente de sélection de caractéristiques [5] et de montrer l'impact du choix de l'espace de couleur pour l'extraction des caractéristiques. Le reste de ce chapitre contient la présentation de l'ensemble de données utilisé pour les tests, la sélection de la région d'intérêt, l'extraction des caractéristiques, la sélection des caractéristiques, la classification. Ensuite, nous présentons les résultats, notre discussion et nos perspectives.

6.2 L'ensemble de donnée : Red Lesion Dataset

Nous avons utilisé l'ensemble de donnée collecté par Paulo et al [108] et disponible publiquement [109]. Il est composé de deux ensembles de données, nommés "Set 1" et "Set 2"

6.2.1 L'ensemble *Set*1

C'est un ensemble de données avec des images aussi diverses que possible provenant de différentes caméras, telles que MiroCam, Pill-Cam SB1, SB2 et SB3 et avec différentes lésions rouges, telles que des angiectasies, des angiodysplasies, du saignement et autres. Il comporte 3295 images dont 1131 présentent des lésions. Toutes les lésions ont été annotées manuellement comme $sang/non-sang$. Les images ont des résolutions de $320x320$.

6.2.2 L'ensemble *Set*2

C'est un ensemble de données avec une séquence de 600 images provenant d'une vidéo PillCam SB3 pour obtenir une évaluation du modèle plus proche de la réalité clinique. L'ensemble contient 439 images avec des lésions rouges, chacune étiquetée manuellement comme $sang/non-sang$ sur la base du jugement humain. Toutes les images ont des résolutions de $320x320$

6.3 Prétraitement des images

Le prétraitement de l'image considéré dans notre étude est la sélection de la région d'intérêt dans les images endoscopiques et le rehaussement de la qualité d'image.

6.3.1 Sélection de la région d'intérêt ROI

Les études dans [110, 111] ont montré que l'extraction de caractéristiques à partir d'images endoscopiques complètes reflète la contamination visuelle présentée dans l'image. En effet, les images WCEs sont obscurcies par le grand fond noir et les bordures. Une illustration de ce facteur est présentée dans **la figure 6.2**.

Pour y remédier, nous délimitons un cercle central à l'intérieur de l'image entière comme ROI, en préservant l'information principale de l'image. Pour cela, une matrice avec des valeurs nulles (pour la région noire) et une (pour la région blanche) est définie. La taille de cette matrice de masque est la même que celle des images de notre base de données. Cette matrice de masque est présentée dans **la figure 6.3** 4. Les ROIs extraits sont satisfaisants (la région blanche) car ils démontrent les principales caractéristiques des images WCE. Dans la suite, nous considérons les images ROI, au lieu de l'image entière. De même, l'extraction des caractéristiques est effectuée sur la région d'intérêt de chaque image.

FIGURE 6.2 – Le fond noir d'une image endoscopique

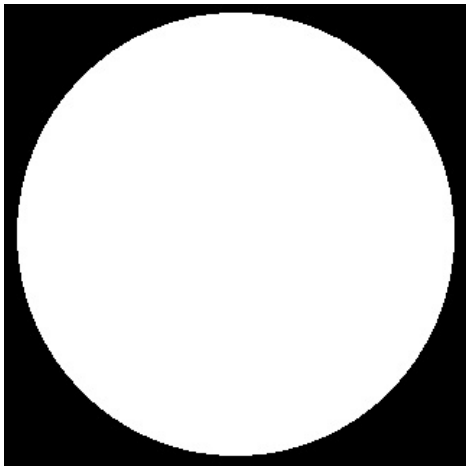

FIGURE 6.3 – Le masque d'image utilisé pour extraire la région d'interêt

6.3.2 Rehaussement de la qualité des images

Le rehaussement ou l'amélioration de la qualité des images [112] est une étape très basique à réaliser mais très importante dans le traitement des images. Cette étape permet de mettre en évidence les données clés tout en compressant ou en supprimant le contenu non informatif dans une image.

Pour améliorer la qualité des images endoscopiques, un débruitage par ondelettes utilisant trois niveaux de décomposition est effectué. Comme recommandé dans les travaux récents [113, 114, 115], nous avons appliqué la méthode de seuillage "doux" (soft thresholding) et avons également utilisé l'ondelette $db2$ pour supprimer le signal indésirable ou le bruit afin de renforcer l'information de saignement pour les séquences d'images WCE.

6.4 Extraction des caractéristiques

Le saignement dans les images WCEs se distingue principalement par sa couleur. De plus, le choix de l'espace de couleur et des caractéristiques est la partie la plus importante qui influence les performances de la classification. Par conséquent, à cette étape, nous extrayons 144 caractéristiques statistiques des ROIs extraites de tous les canaux de différents espaces colorimétriques, y compris les espaces colorimétriques *RGB*, *YCbCr*, *HSI* et *Lab*. Les caractéristiques extraites sont les suivantes : Après l'analyse des résultats de classification, nous pouvons conclure les

f1	Minimum
f2	Maximum
f3	Moyenne
f4	Médiane
f5	Variance
f6	Ecart-type
f7	Contraste
f8	Mode
f9	Inclination
f10	Energie
f11	Entropy
f12	Kurtosis

meilleures caractéristiques et le meilleur espace de couleur pour représenter le saignement dans les images endoscopiques.

6.5 Sélection des caractéristiques

Dans cette partie, nous avons appliqué directement la méthode proposée dans le chapitre 4 $Eigen\_FS$. Cependant, nous avons proposé un nouveau critère afin d'estimer la performance des caractéristiques. En combinant à la fois la redondance et la pertinence dans une seule formule.

Cette formule est basée sur l'information mutuelle normalisée et le score de Fisher. L'information mutuelle et le score de Fisher ont déjà été présentés dans le chapitre 4. Nous allons commencer par introduire l'information mutuelle normalisée.

6.5.1 Information mutuelle normalisée

L'information mutuelle normalisée *NMI* [116] de deux variables aléatoires x et y est définie en fonction de leur information mutuelle $MI(x,y)$, normalisée par l'entropie minimale des deux variables $min(H(x),H(y))$

$$NMI(x,y) = \frac{MI(x,y)}{min(H(x),H(y))} \tag{6.1}$$

6.5.2 Nouveau critère d'évaluation des caractéristiques

Soit *Eva* le nouveau critère d'évaluation des caractéristiques qui prend en considération à la fois la pertinence des caractéristiques et la redondance.

$$Eva_i = \alpha_1 NMI_i + \alpha_2 Fisher_i \tag{6.2}$$

Tel que α_k est le coefficient de mixage (mixing coefficient), $0 \leq \alpha_k \leq 1$ et $\sum \alpha_k = 1$. Les valeurs des coefficients α_k ont été estimées au cours des expériences par validation croisée sur l'ensemble d'apprentissage pour la tâche de classification. En résumé, le score Eva_i indique dans quelle mesure une caractéristique n'est pas redondante (critère de Fisher) et pertinente (information mutuelle normalisée).

6.5.3 Remplissage de la matrice de performance

En utilisant le nouveau critère d'évaluation, le remplissage de la matrice de performance définie dans le chapitre 4 devient comme suit :

$$M_{i,j} = \begin{cases} Eva_{f_i} & \text{si } i = j \\ (Eva_{f_i})(Eva_{f_j}) & \text{sinon.} \end{cases} \tag{6.3}$$

Ainsi, une grande valeur de $\sum M_{i,j}$ implique un bon sous-ensemble. Il convient de noter que la formule ci-dessus n'est qu'une des nombreuses alternatives possibles pour calculer la valeur des caractéristiques f_i et f_j prises ensemble.

La suite de la sélection se fait suivant l'algorithme 4 (chapitre 4) qui résout le problème décrit par la formule (4.13).

6.6 Détection des images des parties saignantes

La partie de la détection des images qui représentent du saignement se fait avec le classifieur SVM qui a été introduit dans le chapitre 2.

Un SVM construit un classifieur binaire à partir d'un ensemble de modèles étiquetés (ensemble d'apprentissage). Soit l'ensemble d'apprentissage suivant :

$$(x_i, y_i) \in R^N - 1, +1, i = 1...m$$

. Le but est de sélectionner une fonction $f : R^N \longrightarrow \pm 1$ parmi une classe de fonctions donnée, de telle sorte que f classifie correctement les données de test (x, y). Sur la base de l'hypothèse ci-dessus, l'algorithme SVM est capable de trouver un hyperplan défini par l'équation :

$$\omega\Phi(x) + \omega_0 = 0$$

Telle que la marge de séparation soit maximisée, où $\Phi(x)$ est une cartographie non linéaire de l'espace d'entrée à l'espace des caractéristiques et ω_0 est un scalaire qui peut être estimé à partir de la condition complémentaire de Karush-Kuhn-Tucker. Il est démontré que pour l'hyperplan de marge maximale (Vapnik, 1995) [47] :

$$\omega = \sum_{i=1}^{m} \lambda_i y_i \Phi^T(x)$$

où λ_i sont les multiplicateurs de Lagrange qui peuvent être estimés à travers la maximisation de :

$$L_D = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j K(x_i, x_j$$

Tel que $\sum_{i=1}^{m} \lambda_i y_i = 0$ et $0 \leq \lambda_i \leq c$. $K(x_i, x_j$ est appelée fonction de noyau et est définie comme le produit interne $K(x_i, x_j) = \omega^T(x_i)\omega(x_j)$. Les fonctions "Linear", "polynomial", "Gaussian", "Radial basis" et la fonction "sigmoid" sont parmi les fonctions les plus utilisées comme noyaux SVM. Étant donné un vecteur d'entrée de test x, le SVM formé produit une sortie qui correspond à l'étiquette de classe, à savoir :

$$s = sign\{\sum_{i=1}^{m} \lambda_i y_i K(x_i, x_j) + \omega_0\}$$

Dans nos expériences, nous avons analysé les données en utilisant trois types de noyaux, à savoir "Linear", "Polynomial" et "Radial Basis Function (RBF)", afin de déterminer le meilleur noyau pour cette classification.

6.7 Protocol expérimental et résultats

Le protocole utilisé est le suivant :

— Itération de plusieurs expériences, en fonction du nombre de caractéristiques sélectionnées λ. L'expérience avec l'ensemble de données avant d'appliquer la sélection des caractéristiques est référencée par "baseline".

— Toutes les expériences ont été réalisées par la méthode de validation croisée 5×2.

— Seule la moyenne des 5 valeurs estimées a été rapportée.

— 40% et 60% de l'ensemble des images représentant des saignements ont été consacrés pour l'ensemble d'apprentissage "train" et l'ensemble de "test" respectivement, à chaque itération de la validation croisée.

— 40% et 60% de l'ensemble des images saines ont été consacrés pour l'ensemble d'apprentissage "train" et l'ensemble de "test" respectivement, à chaque itération de la validation croisée.

— La classification SVM a été itérée avec trois fonctions de noyaux "Linear", "Polynomial" et "RBF".

— Nous validons les performances par l'analyse de l'exactitude "Acc", la sensitivité "Sn" et la spécificité "Sp".

Toutes ces étapes ont été déroulées pour plusieurs valeurs de α_1 et α_2 du critère d'évaluation exprimé par la formule (6.2). Nous avons fait le constat suivant :

— Les meilleurs résultats sont obtenus pour $\alpha_1 = 0.62$ et $\alpha_2 = 0.38$.

— Les meilleurs résultats sont obtenus avec $\alpha_1 \geq 0.55$.

— Les résultats avec $\alpha_1 \geq 0.55$ sont trés rapprochés.

Par conséquent, nous concluons que dans le cas de notre ensemble de données, le critère de la pertinence est un peu plus important que le critère de la redondance. Les résultats présentés dans ce qui suit sont les résultats des tests avec $\alpha_1 = 0.62$ et $\alpha_2 = 0.38$.

6.7.1 Résultat de la classification SVM

Le tableau 6.1 présente les résultats de la classification SVM, en considérant 3 fonctions de noyau. La "moyenne" fait référence à la moyenne des valeurs en considérant plusieurs nombres de caractéristiques sélectionnées, allant de 1 jusqu'à 125, "max" fait référence à la meilleure valeur obtenue et "baseline" fait référence à la classification des données brut, avant d'appliquer la sélection des caractéristiques. **La figure 6.4** est une illustration du **tableau 6.1**.

Les résultats présentés dans **le tableau 6.1** montrent de bonnes performances en général de la détection. Nous remarquons que dans tous les cas les performances après la sélection des caractéristiques ont été considérablement améliorées. Spécialement dans le cas du noyau "RBF". Dans la section suivante, nous allons détailler les résultats de la classification avec le noyau "RBF" pour conclure les meilleures caractéristiques pour détecter les saignements dans les images endoscopiques.

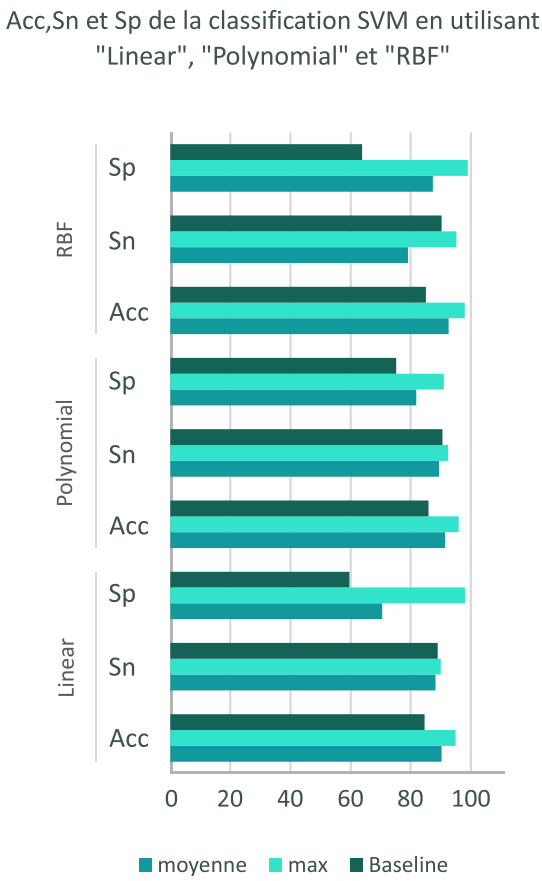

FIGURE 6.4 – Acc,Sn et Sp de la classification SVM en utilisant les fonctions de noyau "Linear", "Polynomial" et "RBF

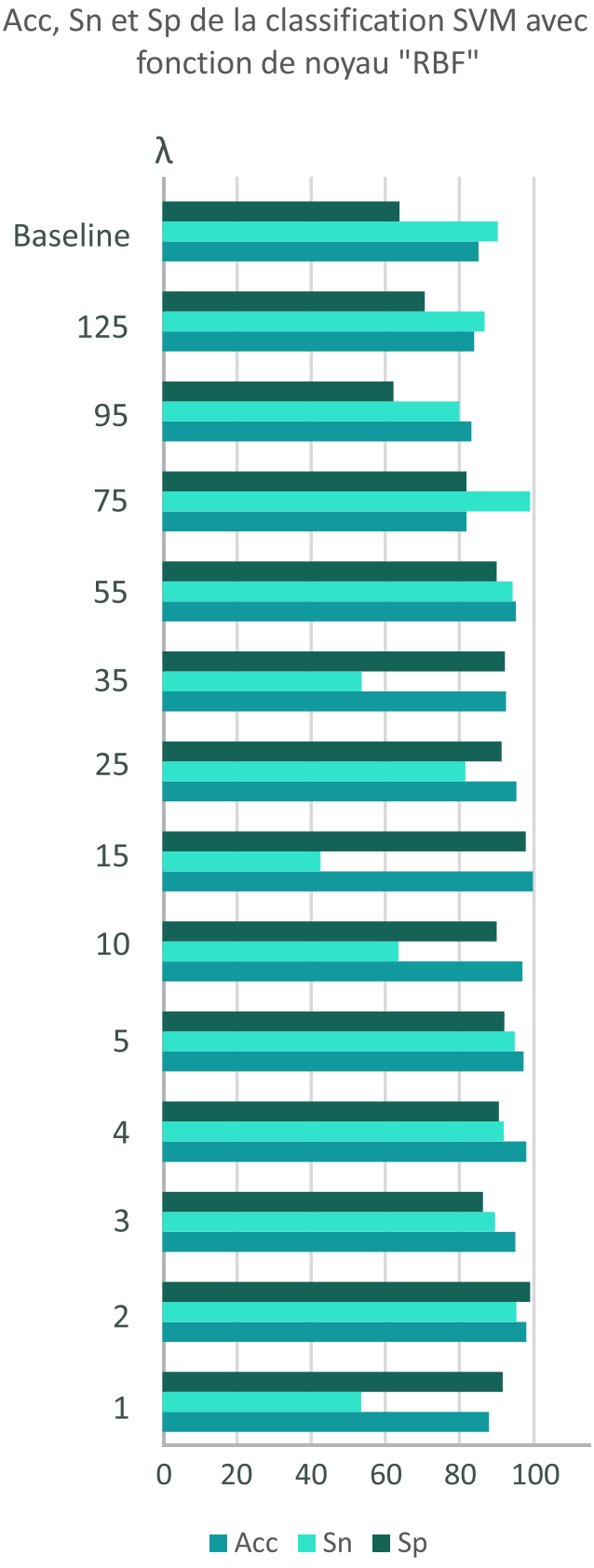

FIGURE 6.5 – Accuracy,Sensitivity et Specificity de la classification SVM en utilisant la fonction de noyau "RBF"

TABLE 6.1 – Acc,Sn et Sp de la classification SVM en utilisant les fonctions de noyau "Linear", "Polynomial" et "RBF"

Noyau	metrique	moyenne	max	Baseline
Linear	Acc	90,3	95	84,6
	Sn	88,2	90	89
	Sp	70,6	98,2	59,6
Polynomial	Acc	91,5	96	86
	Sn	89,4	92,5	90,6
	Sp	81,9	91	75,2
RBF	Acc	92,69	**98**	85,1
	Sn	79,17	**95,3**	90,3
	Sp	87,38	**99,1**	63,9

TABLE 6.2 – Acc,Sn et Sp de la classification SVM en utilisant la fonction de noyau "RBF"

λ	1	2	3	4	5	10	15	25	35	55	75	95	Baseline
Acc	88	**98**	95	98	97,2	96,9	99,8	95,3	92,5	95,2	81,9	83,2	85,1
Sn	53,5	**95,3**	89,5	92	94,9	63,6	42,6	81,6	53,6	94,3	99	80	90,3
Sp	91,7	**99,1**	86,3	90,6	92,1	90	97,9	91,3	92,3	90	81,9	62,2	63,9

6.7.2 Résultat de la classification SVM : noyau "RBF"

Les résultats mentionnés dans **le tableau 6.2** montrent que la sélection des caractéristiques par *Eigen_FS* améliore considérablement les performances de la classification, donc de la détection des saignements dans les images endoscopiques. Les meilleurs performances ont été observées en sélectionnant uniquement 2 caractéristiques parmi 144, avec une exactitude de **98%**, une sensibilité de **95,3%** et une spécificité de **99,1%** contre une exactitude, une sensibilité et une spécificité de $85,1\%$, $90,3\%$ et $63,9\%$ respectivement, avant la sélection des caractéristiques. La sélection de caractéristiques avec une "$Sn = 95,3\%$" est un modèle prometteur pour le domaine médical, en particulier dans le cadre de la détection des saignements dans les images endoscopiques, puisque plus la sensibilité est proche de 100%, moins il y a d'erreurs de détection des sujets suspects (faux négatifs). De plus, notre modèle atteint une "$Sp = 99,1\%$", ceci est très intéressant aussi, puisque plus la spécificité est proche de 100%, moins il y a de faux positifs, ce qui éviterait des coûts supplémentaires. Les deux meilleures caractéristiques sont la moyenne issue du canal H et la moyenne issue du canal S de l'espace de couleur HSI. Nous pouvons directement conclure que l'espace HSI est le plus représentatif du saignement dans les images endoscopiques. L'espace colorimétrique HSI présente de nombreuses similitudes avec la perception de l'œil humain.

Dans ce qui suit, nous allons extraire, les caractéristiques seulement de l'espace de couleur HSI pour voir le comportement des autres caractéristiques.

6.7.3 Résultat de la classification SVM en considérant l'espace de couleur HSI

Dans la section précédente, nous avons conclu que l'espace de couleur HSI est le plus représentatif du saignement dans les images endoscopiques. En effet, l'espace colorimétrique HSI présente de nombreuses similitudes avec la perception de l'œil humain. Ici, le signal d'image donné est dissous dans la partie chrominance H et S (Hue et Saturation) et la partie luminance I (Intensity) pour l'espace HSI, qui est similaire au système de perception visuelle humaine [117].

Nous allons donc faire des tests avec un ensemble de caractéristiques extraites de cet espace de couleur. **Le tableau 6.3** représente l'exactitude, la sensibilité et la spécificité de la détection des images qui contient du saignement, avec un vecteur de caractéristiques extraites de l'espace de couleur HSI. λ représente le nombre de caractéristiques sélectionnées avec la méthode $Eigen\_FS$. L'ensemble CS/Can contient des couples de caractéristique sélectionnée (CS) et de quel canal de couleur provient cette caractéristique (Can).

TABLE 6.3 – Résultat de la classification SVM avec la sélection des caractéristiques extraites de l'espace de couleur HSI

λ	1	2	3	4	5	6	7	8	9	10	Baseline
Acc	88,1	98	95	98	97,2	95,6	88,9	62,3	62,3	55,9	93,2
Sn	60,3	95,3	89,5	92	94,9	85,3	86,6	90	56	81,6	42
Sp	81	99,1	86,3	90,6	92,1	83	83	82,2	75,3	60	85,3
{CS/Can}	{f3/S}	{f3/H, f3/S }	{f3/H, f3/S, f6/S}	{f3/H, f6/S, f12/S, f6/I}	{f3/H, f12/H, f11/S, F2/S, f6/I}	{f3/H, f12/H, f6/I, f3/S, f12/S, f10/H}	{f3/H, f1/H, f12/S, F11/I, F11/H, f6/H, f10/H}	{f3/H, f1/H, F5/S, F4/L, F2/H, F3/I, Ff/I, F6/S }	{f3/H, f12/S, F5/H, f6/I, f12/I, f1/H, f12/S, f10/S, f11/I}	{f3/H, f1/I, f12/S, f2/I, f6/H, f1/H, f11/S, f6/S, f12/H, f11/H}	Les 36 caractéristiquess extraites

Les résultats présentés dans **le tableau 6.3** montrent les performances de la détection du saignement, dans l'espace de couleur *HSI*, nous remarquons que notre approche adoptée améliore significativement les performance de l'approche de base, c'est à dire avant la sélection des caractéristiques. Nous remarquons aussi que les combinaisons (caractéristiques sélectionnées/ canal) qui reviennent le plus c'est (f3/H) avec 10 apparitions et (f12/S) avec 6 apparitions, qui sont respectivement la moyenne extraite du canal H et le kurtosis extrait du canal S, ceci pour les combinaisons de 1 à 11 meilleures caractéristiques. Le canal de couleur qui revient le plus c'est le canal H et la caractéristique qui revient le plus c'est la $f3$, ceci est toujours pour les combinaisons de 1 à 11 meilleures caractéristiques. Nous remarquons aussi que la caractéristique $f4$ est apparue une seule fois, dans une combinaison de 8 caracteristiques. De plus, les caractéristique $\{f7, f8, f9\}$ qui sont respectivement la médiane, le contraste, le mode et l'inclination ne sont jamais sélectionnées (dans le cas de combinaison de taille 1 à 11). Toutes les combinaisons d'apparition sont montrées dans **le tableau 6.4**.

TABLE 6.4 – Résumé des caractéristiques sélectionnées en fonction du canal colorimétrique

CS/Can	H	S	I
f1	5	0	2
f2	1	1	2
f3	10	4	1
f4	0	0	1
f5	1	1	0
f6	3	5	4
f7	0	0	0
f8	0	0	0
f9	0	0	0
f10	3	1	0
f11	2	3	2
f12	4	6	1

Pour valider les performances de notre approche de sélection $Eigen\_FS$, nous avons fait des tests de classification avec des combinaisons de caractéristiques. Il est très difficile et fastidieux de faire les tests avec toutes les combinaisons, nous avons donc limité nos expériences à des combinaisons avec un nombre allant de 1 caractéristique à 5 caractéristiques. De plus, les premiers tests qui ont montrés l'inutilité des caractéristiques $\{f4, f7, f8, f9\}$, nous ont permis d'éliminer ces 4 caractéristiques. Nous devons observer les caractéristiques de chaque combinaison qui donne la meilleur accuracy. Ceci est discuté dans la section suivante.

6.7.4 Résultats de la classification SVM des vecteurs de caractéristiques manuelles

Les résultats expérimentaux obtenus en utilisant la méthode manuelle sont montrés dans **le tableau 6.5**. Toutes les combinaisons de 1 à 5 caractéristiques ont été considérées dans nos

tests. Cependant, seules les combinaisons ayant les meilleurs résultats ont été mentionnées dans le tableau. Les combinaisons sont référencées par CS/Can dans le tableau. L'accuracy, la sensibilité et la spécificité ont été calculées dans le cadre de ces expériences. Dans le même tableau, nous avons mentionné les résultats obtenus par une sélection automatiques en utilisant notre approche de sélection $Eigen\_FS$. La colonne *n* du tableau représente la taille de la combinaison de caractéristiques.

TABLE 6.5 – Acc, Sn et Sp de la classification SVM en utilisant des combinaisons manuelles de caractéristiques et la sélection $Eigen\_FS$

n	Sélection	CS\Can	Acc	Sn	Sp
1	$Eigen\_FS$	{f3/S}	88,1	60,3	81
	Manuelle	{f3/S}	88,1	60,3	81
2	$Eigen\_FS$	{f3/H, f3/S}	98	95,3	99,1
	Manuelle	{f3/H, f3/S}	98	95,3	99,1
3	$Eigen\_FS$	{f3/H,f6/H,f6/S}	95	89,5	**86,3**
	Manuelle	**{f3/H,f3/S,f6/S}**	**96,2**	**91,2**	85,1
4	$Eigen\_FS$	{f3/H,f6/S,f12/S, f6/I}	98	92	90,6
	Manuelle	{f3/H,f6/S,f12/S, f6/I}	98	92	90,6
5	$Eigen\_FS$	{f3/H,f12/H,f11/S,F2/S,f6/I}	97,2	94,9	92,1
	Manuelle	{f3/H,f12/H,f11/S,F2/S,f6/I}	97,2	94,9	92,1

Il est essentiel de mentionner que pour une comparaison équitable, nous avons déroulé les expériences avec les mêmes sous-ensembles de données divisés par la méthode de la validation croisée.

Les résultats mentionnés dans **le tableau 6.5** montrent que notre approche de sélection est efficace dans la plupart des cas. Tel que, les combinaisons sélectionnées par notre approche sont les meilleures trouvées dans l'approches manuelle dans les cas suivants : combinaison d'une caractéristique, combinaison de deux, de quatre et de cinq caractéristiques. La combinaison de trois caractéristiques a été faussement sélectionnée par notre approche, tel que $Eigen\_FS$ à sélectionné la combinaison suivante : $\{f3/H, f3/S, f6/S\}$ par contre manuellement la meilleure combinaison de caractéristiques est la suivante : $\{f3/H, f6/H, f6/S\}$. Il est remarquable, que cette combinaison est meilleure que la combinaison sélectionnée par $Eigen\_FS$ en termes d'accuracy et de sensibilité, mais pas en termes de spécificité. Nous constatons que malgré l'excellence du nouveau critères d'évaluation de caractéristiques, la relation par paire des caractéristiques est trop présente, puisque la caractéristique $\{f3/S\}$ est absente de la combinaison manuelle, mais aussi le couple de caractéristique $\{f3/H, f3/S\}$ est présent dans la combinaison sélectionnée et aussi c'est le couple qui donne les meilleurs performances du système, non seulement dans l'approche manuelle mais aussi dans l'approche qui utilise la sélection des caractéristiques $Eigen\_FS$. Ce n'est qu'une hypothèse qui ouvre de nouvelles perspectives de recherches et de tests exhaustifs.

6.8 Conclusion et perspectives

Dans ce chapitre, nous avons principalement appliqué la nouvelle méthode de sélection de caractéristiques proposé dans le domaine du traitement d'images médicale. Plus précisément, sur la classification des images endoscopiques qui contiennent du saignement. Les images sont issues de la base de données "Red Lesion Dataset". Nous avons commencé par spécifier la région d'intérêt dans les images endoscopiques et par améliorer la qualité des images, avant de procéder à l'extraction des caractéristiques. Notre contribution principale, à part le test de cette nouvelle méthode de sélection est la suivante :

— Nous avons proposé un nouveau critère d'évaluation de caractéristiques, qui prend à la fois en considération la pertinence et la redondance des caractéristiques.

— Nous avons fait l'extraction des caractéristiques à partir de plusieurs espaces de couleurs pour savoir le meilleur qui représente le saignement dans les images endoscopiques.

— Nous avons fait des tests de détection avec des combinaisons de caractéristiques choisies manuellement, pour valider notre approche de sélection.

— Grâce aux tests, nous avons pu conclure les meilleures caractéristiques qui représente le saignement dans les images endoscopiques.

Après avoir exécuté une série de test, nous avons conclu que la classification SVM en utilisant la fonction de noyau "RBF" est la meilleure dans le cadre de cette étude. Après avoir analysé les vecteurs de caractéristiques et résultats de la classification, nous avons conclu que dans le cadre de la détection des saignements dans les images endoscopiques, l'espace colorimétrique le plus représentatif est l'espace HSI, qui présente de nombreuses similitudes avec la perception de l'œil humain. Nous avons ensuite exécuté plusieurs expériences de classification, en utilisant uniquement des caractéristiques sélectionnées de l'espace HSI. Nous avons conclu que les meilleures caractéristiques pour représenter le saignement dans les images endoscopiques sont la moyenne de la teinte et de la saturation $f3/H$ et $f3/S$ respectivement, cela a donné des taux d'accuracy, sensibilité et spécificité de 98%, $95,3\%$ et $99,1\%$ respectivement. Une accuracy de 98% a été atteinte aussi avec une combinaison de 4 caractéristiques, mais la sensibilité et la spécificité sont de 92% et $90,6\%$ respectivement. Les tests de toutes les combinaisons manuellement a montré les performances de notre approche de sélection $Eigen\_FS$. Cependant, nous posons l'hypothèse que la relation par pair est dominante dans le critère d'évaluation proposé dans cette étude. Notre perspective est de proposer d'autres critères et faire une comparaison. En conclusion, l'approche proposée est prometteuse pour la distinction entre les saignements et les images des parties saines. Cette méthode peut également être testée pour d'autres anomalies dans les images endoscopiques.

Conclusion Générale

Conclusion

En raison de la croissance explosive des technologies numériques, de plus en plus de données visuelles sont générées et stockées. De nos jours, les données visuelles sont aussi courantes que les données textuelles. Il y a un besoin urgent d'un outil efficace qui permet de trouver les informations visuelles à la demande. C'est pour cette raison, au cours des dernières décennies, l'un des domaines de recherche les plus prolifiques est l'analyse d'images. Différentes techniques ont été développées pour analyser automatiquement les images. L'analyse se fait principalement à partir des caractéristiques visuelles extraites, comme la couleur, la texture, la forme. Le défi est d'identifier la relation entre ces caractéristiques de données et un certain point final, par exemple la classe cible pour une tâche de classification. Dans la plupart des ensembles de données, seules quelques caractéristiques sont pertinentes et contribuent à déterminer le point final. Les autres caractéristiques contribuent à la dimensionnalité globale de l'espace du problème, ce qui implique une mémoire importante pour stocker toutes les caractéristiques, un temps de traitement important pour obtenir le résultat souhaité, ou des résultats biaisés à cause des caractéristiques bruitées. Par conséquent, le prétraitement des données est d'une grande importance. La sélection d'un sous-ensemble de caractéristiques est considérée comme une solution appropriée pour résoudre les problèmes ci-dessus, et est devenue une composante primordiale du processus d'apprentissage automatique. Au cours de cette thèse, nous nous sommes intéressées à la sélection des caractéristiques, nous avons proposé une nouvelle approche générique qui peut être appliquée dans différents domaine. Pour démontrer l'efficacité de notre approche, nous l'avons appliqué au profit des système d'aide au diagnostic médical pour la détection des anomalies dans le tracus gastrointestinal. Ce qui a motivé notre choix d'application est les statistiques citées ci-dessous. Le nombre de cas de cancers gastrointestinaux est en croissante exponnentielle, représentant à ce jour 2,8 millions de nouveaux cas et 1,8 million de décès par an dans le monde, légèrement plus élevée chez les hommes quechez les femmes. Lorsque les cellules saines de la paroi du tractus gastro-intestinal (GI) se développent horsdu contrôle normal, quelque chose d'anormal se produit, comme un saignement rectal, une inflammation ou la formation d'une masse. Lorsque les anomalies sont détectées à unstade précoce, la guérison est souvent possible. Par conséquent, pour sauver la vie dupatient, nous devons detecter les cellules affectées à un stade précoce. En 2000, l'endoscopie par cap-sule sans fil

"Wireless Capsule Endoscopy" (WCE) a été développée dans le cadre du diagnostic des anomalies gastrointestinales. Cette technique est utilisée de façon indolore, mais le nombre d'images résultantes de cette opération est trés important, alors que seulement 5% environ de ces images peuvent être anormales chez les patients souffrant de maladies gastro-intestinales. Dans ce cas, la détection des anomalies est une tâche fastidieuse pour le médecin qui doit analyser visuellement toutes les images. Par conséquent, un système assisté par ordinateur est envisagé pour résoudre ce problème de temps.L'amélioration du diagnostic et le fardeau économique médical sont actuellement abordés,puisque les chercheurs se sont concentrés sur le développement d'outils de diagnostic assisté par ordinateur (DAO) basés sur le concept de l'apprentissage automatique "Machine Learning"(ML). La question clé de recherche qui motive cette thèse c'est comment améliorer la précision des outils DAO ? La réponse est, en améliorant le module final qui consiste à classer les images selon le problème à résoudre. Une autre question de recherche se pose, c'est comment améliorer la précision de la classification ? Une bonne représentation de données implique un bon apprentissage et donc une bonne précision de classification. Pour la préparation des données à la classification, un image passe par plusieurs étapes :

— Le prétraitement, ceci inclut à la fois l'amélioration de la qualité d'image et la sélection de la région d'interet.

— l'extraction des caractéristiques, c'est l'ensemble de méthodes qui traitent toute transformation ou combinaison de données d'entrée en un vecteur de caractéristiques.

— La sélection des caractéristiques, ceci a pour but d'éliminer les caractéristiques redondantes et non pertinente ou bruyante.

Le prétraitement envisagé au cours de cette étude et le débruitage par ondelette et la sélection de la région d'interet. L'extraction de caractéristiques a été faite à partir de plusieurs éspace de couleurs, puisque l'éspace de couleur qui représente le mieux un problème n'est pas connu à priori. Nous avons fait l'extraction de 144 caractéristiques statistiques, à partir de 4 différents espaces de couleurs, à savoir *RGB*, *YCbCr*, *HSI* et *Lab*. Pour la sélection des caractéristiques, au cours de cette thèse, nous nous sommes interessés au problème de la sélection des caractéristiques visuelles pour la classification des images endoscopiques, dans le but de la détection des saignement. Nous avons commencé par définir en détail le problème de la sélection des caractéristiques en général. La sélection des caractéristiques est primordial dans le processus de l'apprentissage automatique, qui permet une meilleur comprehension des données, une réduction de la complexité du problème, la réduction de l'espace de stockage et de temps de résolution. Les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection sont classées en trois grandes catégories principales : "filter", "wrapper" et "embedded" et elles sont basées sur le clustering, la classification ou l'optimisation. Nous avons ensuite passé en revu quelques méthodes de classification selon le domaine d'application, et nous nous sommes interessées particulièrement à la sélection des caractéristiques pour les problèmes de

classification. Nous avons introduit en détail les différentes techniques d'apprentissage automatique, à savoir l'apprentissage supervisé, semi-supervisé et non supervisé. Ensuite, nous avons introduit la classification des images en se basant sur les réseaux de neurones artificiels profond, nous avons terminé cette section par la présentation d'un nouveau modèle profond pour la classification des images endoscopiques en se basant sur la technique d'apprentissage par transfert. Nous avons ensuite proposé une nouvelle approche de sélection de caractéristiques basée sur le calcul des valeurs propres avec des contraintes linéaire $Eigen\_FS$. Nous avons résolu notre problème avec un algorithme proposé par Xu et al [89] , cet algorithme est éfficace et converge vers la solution optimale globale. Dans un premier temps, nous avons testé et validé notre approche avec des données réelles, provenant du réferentiel *UCI Machine Learning*. Nous avons commencé par l'approximation de notre problème à un problème de coupe maximal, ensuite nous l'avons réduit à un problème de calcul de valeurs propres avec des contraintes lineaires, pour respecter le nombres total de caractéristiques à sélectionnées. Les caractéristiques de notre contribution est qu'elle prend des avantage des approches de classement basées sur l'optimisation, tout en remédiant à leur limites. Nous avons Notre contrinution aussi dans le cadre de la detection du saignement dans les images endoscopiques est la proposition d'un nouveau critère d'évaluation des caractéristiques. Nous avons aussi fait une analyse des caractéristiques sélectionnées et nous avons conclu que l'espace de couleur le plus représentatif du saignement est l'espace HSI qui présente de nombreuses similitudes avec la perceptionde l'œil humain. Comme complément de cette étude d'éspace de couleur, nous avons opéré la sélection de caractéristiques, uniquement à partir des caractéristiques extraites de l'espace de couleur HSI et nous avons conclu que les meilleures caractéristiques qui représente le saignement dans les images endoscopiques par capsules sont la moyenne issue du canal H et la moyenne issue du canal S. Pour valider les performances de notre approches, nous avons effectuer des tests manuelle en parallèle de la sélection, c'est à dire, nous avons fait la classification avec des combinaisons de caractéristiques sélectionnées de façon exaustive et manuelle. Nous avons tiré deux conclusion :

— Le schéma adopté pour la detection de saignement est performant et optimal dans la plus part des cas.

— Dans le cas ou la solution n'est pas optimal, d'aprés notre analyse, nous supposons que la relation des caractéristiques par paires est dominante dans le nouveau critère d'évaluation de caractéristiques proposé au cours de cette thèse.

Une autre contribution est le test de la classification SVM avec trois différentes fonctions de noyau, à savoir : "Linear", "Plynomial" et "RBF". Nous avons conclu que la meilleure fonction d enoyau au cours de cette étude est la fonction "RBF".

Points forts et points faibles

Les avantages de notre approche de sélection sont les suivants :

— Elle résoud le problème du hasard des approches d'optimisation. Tel que, le point de départ de l'optimisation n'est pas aléatoire mais c'est le vecteur des valeurs propre de la matrice des performances.

— Elle prend en compte à la fois la pertinence des caractéristiques et la redondance, ceci est assuré par le nouveau critère de l'évaluation des caractéristiques.

— Elle prend en compte la relation par paire des caractéristiques, contraitement aux approches de classement qui prennent pas en compte la relation entre les caractéristiques.

— Elle améliore considérablement les performance de la classification des images endoscopique. L'accuracy, la sensibilité et la spécifité aprés la sélection des caractéristiques par $Eigen\_FS$ sont de 98%, $95,3\%$ et $99,1\%$ contre $93,2\%$, 42% et $85,3\%$ réspectivement avant la sélection.

— Les sous-ensembles de caractéristiques sélectionnées est optimal dans la plus part des cas, ceci dans le cadre de la detection du saignement dans les images endoscopiques.

— Les performances de la sélection $Eigen\_FS$ dépassent les performances d'une méthode de pointe, à savoir *mRMR*. Ceci dans le cadre de classification diverses, en utilisant des ensembles de données du référentiel UCI.

Bien que l'efficacité de notre approche de sélection $Eigen\_FS$ et notre schéma complet de detection ont été montré, quelques limites sont à citer :

— La limite majeure c'est que la sélection ne se fait pas automatiquement, mais il faut introduire au préalable le nombre de caractéristiques à sélectionner. Cependant, ça aide à éliminer toutes les caractéristiques inutiles.

— Nous posons l'hypothèse que la relation par paire formulée dans le critère d'évaluation des caractéristiques proposé est dominante.

Perspectives

Des orientations intéressantes se dégagent de l'analyse et la discussion des résultats et pourront être approfondies lors de recherches ultérieures, à savoir :

— Bien que l'introduction du nombre de caractéristiques à sélectionner nous a permi de déterminer les meilleures caractéristiques, les caractéristiques inutiles, les canaux de couleur les plus interessants mais une sélection automatique serait plus interessante, pour résoudre le problème de la plupart des méthodes de la littérature. Cela pourrait se faire, par éliminer la contrainte linéaire de notre problème à optimiser et en rendant flexible l'algorithme de la solution "Projected Power".

— La proposition d'un nouveau critère d'évaluation serait une bonne alternative pour confirmer ou pas l'hypothèse posée concernant le critère déja proposé.

— Appliquer le même schéma pour le problème de multiclassification proposé dans le chapitre 2, en utilisant le même ensemble de données *Kvasir*.

Nos perspectives sont motivés par l'efficacité approuvé de notre schéma présenté dans cette thèse.

Bibliographie

[1] Amina Benkessirat and Nadjia Benblidia. Fundamentals of feature selection : an overview and comparison. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2019.

[2] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2) :1, 2015.

[3] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar) :1157–1182, 2003.

[4] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 5(4) :83–124, 2011.

[5] Amina Benkessirat and Nadjia Benblidia. A novel feature selection approach based on constrained eigenvalues optimization. *Journal of King Saud University-Computer and Information Sciences*, 2021.

[6] Faten Hussein, Nawwaf Kharma, and Rabab Ward. Genetic algorithms for feature selection and weighting, a review and study. In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pages 1240–1244. IEEE, 2001.

[7] Tofigh Naghibi, Sarah Hoffmann, and Beat Pfister. A semidefinite programming based search strategy for feature selection with mutual information measure. *IEEE transactions on pattern analysis and machine intelligence*, 37(8) :1529–1541, 2014.

[8] Andreas Moser and M Narasimha Murty. On the scalability of genetic algorithms to very large-scale feature selection. In *Workshops on Real-World Applications of Evolutionary Computation*, pages 77–86. Springer, 2000.

[9] Anirudha Majumdar, Georgina Hall, and Amir Ali Ahmadi. A survey of recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *arXiv preprint arXiv :1908.05209*, 2019.

[10] Amina Benkessirat, Nadjia Benblidia, and Azeddine Beghdadi. Gastrointestinal image classification based on vgg16 and transfer learning. In *2021 International Conference on Information Systems and Advanced Technologies (ICISAT)*, pages 1–5. IEEE, 2021.

[11] Muhammad Iqbal, Malik Muneeb Abid, Muhammad Noman Khalid, and Amir Manzoor. Review of feature selection methods for text classification. *International Journal of Advanced Computer Research*, 10(49) :2277–7970, 2020.

[12] Vipin Kumar and Sonajharia Minz. Feature selection : a literature review. *SmartCR*, 4(3) :211–229, 2014.

[13] George H John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier, 1994.

[14] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2) :273–324, 1997.

[15] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2) :245–271, 1997.

[16] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4) :131–156, 1997.

[17] Patrenahalli M. Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Computer Architecture Letters*, 26(09) :917–922, 1977.

[18] Justin Doak. An evaluation of feature selection methods and their application to computer security. *Techninal Report CSE-92-18*, 1992.

[19] P Ravi Kiran Varma, V Valli Kumari, and S Srinivas Kumar. A survey of feature selection techniques in intrusion detection system : A soft computing perspective. In *Progress in computing, analytics and networking*, pages 785–793. Springer, 2018.

[20] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1) :16–28, 2014.

[21] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19) :2507–2517, 2007.

[22] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee, 2015.

[23] Hussein Almuallim and Thomas G Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial intelligence*, 69(1-2) :279–305, 1994.

[24] M Ben-Bassat. Pattern recognition and reduction of dimensionality. *Handbook of Statistics*, 2(1982) :773–910, 1982.

[25] Robert J Schalkoff. Pattern recognition. *Wiley Encyclopedia of Computer Science and Engineering*, 2007.

[26] Mark A Hall. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

[27] Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media, 2012.

[28] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

[29] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35 :99–109, 1943.

[30] E Chandra Blessie and E Karthikeyan. Sigmis : A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, 6(3) :385–394, 2012.

[31] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4) :491–502, 2005.

[32] Manoranjan Dash and Huan Liu. Feature selection for clustering. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 110–121. Springer, 2000.

[33] Jennifer G Dy and Carla E Brodley. Feature subset selection and order identification for unsupervised learning. In *ICML*, pages 247–254. Citeseer, 2000.

[34] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug) :845–889, 2004.

[35] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4) :491–502, 2005.

[36] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4) :491–502, 2005.

[37] Julliano Trindade Pintas, Leandro AF Fernandes, and Ana Cristina Bicharra Garcia. Feature selection methods for text classification : a systematic literature review. *Artificial Intelligence Review*, pages 1–52, 2021.

[38] Thomas Marill and D Green. On the effectiveness of receptors in recognition systems. *IEEE transactions on Information Theory*, 9(1) :11–17, 1963.

[39] A Wayne Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 100(9) :1100–1103, 1971.

[40] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8) :1226–1238, 2005.

[41] Kenji Kira, Larry A Rendell, et al. The feature selection problem : Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

[42] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1) :23–69, 2003.

[43] Patrenahalli M. Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on computers*, 26(09) :917–922, 1977.

[44] Xue-wen Chen. An improved branch and bound algorithm for feature selection. *Pattern Recognition Letters*, 24(12) :1925–1933, 2003.

[45] Maneela Jain and Pushpendra Singh Tomar. Review of image classification methods and techniques. *International journal of engineering research and technology*, 2(8), 2013.

[46] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[47] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1) :1–37, 2008.

[48] Lior Rokach and Oded Maimon. Decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005.

[49] Rung-Ching Chen and Chung-Hsun Hsieh. Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications*, 31(2) :427–435, 2006.

[50] Pádraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, 54(6) :1–25, 2021.

[51] Mohammad Reza Bonyadi, Quang M Tieng, and David C Reutens. Optimization of distributions differences for classification. *IEEE transactions on neural networks and learning systems*, 30(2) :511–523, 2018.

[52] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7 :1–30, 2006.

[53] Jerrold H Zar. *Biostatistical analysis*. Pearson Education India, 1999.

[54] Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6) :571–595, 1980.

[55] Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua. Going deep in medical image analysis : Concepts, methods, challenges, and future directions. *IEEE Access*, 7 :99540–99572, 2019.

[56] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42 :60–88, 2017.

[57] Bilel Sdiri, Faouzi Alaya Cheikh, Kushtrim Dragusha, and Azeddine Beghdadi. Comparative study of endoscopic image enhancement techniques. In *2015 Colour and Visual Computing Symposium (CVCS)*, pages 1–5. IEEE, 2015.

[58] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8) :5455–5516, 2020.

[59] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[60] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25 :1097–1105, 2012.

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[62] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[63] Siyuan Lu, Zhihai Lu, and Yu-Dong Zhang. Pathological brain detection based on alexnet and transfer learning. *Journal of computational science*, 30 :41–47, 2019.

[64] Taruna Agrawal, Rahul Gupta, Saurabh Sahu, and Carol Y Espy-Wilson. Scl-umd at the medico task-mediaeval 2017 : Transfer learning based classification of medical images. In *MediaEval*, 2017.

[65] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir : A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[66] Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, and Muhammad Atif Tahir. Ensemble of texture features for finding abnormalities in the gastro-intestinal tract. In *MediaEval*, 2017.

[67] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. An inception-like cnn architecture for gi disease and anatomical landmark classification. In *MediaEval*, 2017.

[68] Yang Liu, Zhonglei Gu, and William K Cheung. Hkbu at mediaeval 2017 medico : Medical multimedia task. 2017.

[69] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7 :1–26, 2020.

[70] Esin Guldogan and Moncef Gabbouj. Feature selection for content-based image retrieval. *Signal, Image and Video Processing*, 2(3) :241–250, 2008.

[71] Olfa Allani, Nedra Mellouli, Hajer Baazaoui Zghal, Herman Akdag, and Henda Ben Ghézala. A relevant visual feature selection approach for image retrieval. In *VISAPP (2)*, pages 377–384, 2015.

[72] Xiang Sean Zhou, Ira Cohen, Qi Tian, and Thomas S Huang. Feature extraction and selection for image retrieval. In *ACM Multimedia 2000*. Citeseer, 2000.

[73] Khawla Tadist, Said Najah, Nikola S Nikolov, Fatiha Mrabti, and Azeddine Zahi. Feature selection methods and genomic big data : a systematic review. *Journal of Big Data*, 6(1) :1–24, 2019.

[74] Majid Afshar and Hamid Usefi. High-dimensional feature selection for genomic datasets. *Knowledge-Based Systems*, 206 :106370, 2020.

[75] Ioannis Tsamardinos, Giorgos Borboudakis, Pavlos Katsogridakis, Polyvios Pratikakis, and Vassilis Christophides. A greedy feature selection algorithm for big data of high dimensionality. *Machine learning*, 108(2) :149–202, 2019.

[76] Yang Shen, Jintian Xu, Zhiyong Li, Yichen Huang, Ye Yuan, Jixiang Wang, Meng Zhang, Songnian Hu, and Ying Liang. Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia : a cross-sectional study. *Schizophrenia research*, 197 :470–477, 2018.

[77] Zahid Halim, Muhammad Nadeem Yousaf, Muhammad Waqas, Muhammad Suleman, Ghulam Abbas, Masroor Hussain, Iftekhar Ahmad, and Muhammad Hanif. An effective genetic algorithm-based feature selection method for intrusion detection systems. *Computers & Security*, page 102448, 2021.

[78] Vitali Herrera-Semenets, Lázaro Bustio-Martínez, Raudel Hernández-León, and Jan van den Berg. A multi-measure feature selection algorithm for efficacious intrusion detection. *Knowledge-Based Systems*, 227 :107264, 2021.

[79] Firuz Kamalov, Sherif Moussa, Rita Zgheib, and Omar Mashaal. Feature selection for intrusion detection systems. In *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, pages 265–269, 2020.

[80] SB Pooja, RV Siva Balan, M Anisha, MS Muthukumaran, and R Jothikumar. Techniques tanimoto correlated feature selection system and hybridization of clustering and boosting ensemble classification of remote sensed big data for weather forecasting. *Computer Communications*, 151 :266–274, 2020.

[81] Yang Shen, Jintian Xu, Zhiyong Li, Yichen Huang, Ye Yuan, Jixiang Wang, Meng Zhang, Songnian Hu, and Ying Liang. Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia : a cross-sectional study. *Schizophrenia research*, 197 :470–477, 2018.

[82] Yi Zhou, Rui Zhang, Shixin Wang, and Futao Wang. Feature selection method based on high-resolution remote sensing images and the effect of sensitive features on classification accuracy. *Sensors*, 18(7) :2013, 2018.

[83] K Thirumoorthy and K Muneeswaran. Feature selection for text classification using machine learning approaches. *National Academy Science Letters*, pages 1–6, 2021.

[84] Yong Liu, Shenggen Ju, Junfeng Wang, and Chong Su. A new feature selection method for text classification based on independent feature space search. *Mathematical Problems in Engineering*, 2020, 2020.

[85] Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E. Alsaadi. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86 :105836, 2020.

[86] Giorgio Roffo, Umberto Castellani, Alessandro Vinciarelli, Marco Cristani, et al. Infinite feature selection : a graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[87] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4) :537–550, 1994.

[88] Konstantinos Sechidis and Gavin Brown. Simple strategies for semi-supervised feature selection. *Machine Learning*, 107(2) :357–395, 2018.

[89] Linli Xu, Wenye Li, and Dale Schuurmans. Fast normalized cut with linear constraints. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2866–2873. IEEE, 2009.

[90] Arthur Asuncion and David Newman. Uci machine learning repository, 2010.

[91] Vangelis Th Paschos. *Applications of combinatorial optimization*, volume 3. John Wiley & Sons, 2014.

[92] Qingzhi Yang, Yiyong Li, and Pengfei Huang. A novel formulation of the max-cut problem and related algorithm. *Applied Mathematics and Computation*, 371 :124970, 2020.

[93] Clayton W Commander. Maximum cut problem, max-cut. *Encyclopedia of Optimization*, 2, 2009.

[94] Svatopluk Poljak, Franz Rendl, and Henry Wolkowicz. A recipe for semidefinite relaxation for (0, 1)-quadratic programming. *Journal of Global Optimization*, 7(1) :51–73, 1995.

[95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6) :1115–1145, 1995.

[96] Ron Larson. *Elementary linear algebra*. Cengage Learning, 2016.

[97] HongFang Zhou, Yao Zhang, YingJie Zhang, and HongJiang Liu. Feature selection based on conditional mutual information : minimum conditional relevance and minimum conditional redundancy. *Applied Intelligence*, 49(3) :883–896, 2019.

[98] Jens Keuchel, Christoph Schnorr, Christian Schellewald, and Daniel Cremers. Binary partitioning, perceptual grouping, and restoration with semidefinite programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11) :1364–1379, 2003.

[99] David S Watkins. *Fundamentals of matrix computations*. John Wiley & Sons, 2004.

[100] Walter Gander, Gene H Golub, and Urs Von Matt. A constrained eigenvalue problem. *Linear Algebra and its applications*, 114 :815–839, 1989.

[101] Hadjer Ykhlef and Djamel Bouchaffra. An efficient ensemble pruning approach based on simple coalitional games. *Information Fusion*, 34 :28–42, 2017.

[102] Health Quality Ontario et al. Colon capsule endoscopy for the detection of colorectal polyps : an evidence-based analysis. *Ontario health technology assessment series*, 15(14) :1, 2015.

[103] Tariq Rahim, Muhammad Arslan Usman, and Soo Young Shin. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Computerized Medical Imaging and Graphics*, 85 :101767, 2020.

[104] Cancer.Net Editorial Board. Colorectal-cancer : statistics, 2020.

[105] Zahra Amiri, Hamid Hassanpour, and Azeddine Beghdadi. A computer-aided method to detect bleeding frames in capsule endoscopy images. In *2019 8th European Workshop on Visual Information Processing (EUVIP)*, pages 217–221. IEEE, 2019.

[106] Rosdiana Shahril, Sabariah Baharun, and AKM Muzahidul Islam. Pre-processing technique for wireless capsule endoscopy image enhancement. *International Journal of Electrical and Computer Engineering*, 6(4) :1617, 2016.

[107] Amit Kumar Kundu, Shaikh Anowarul Fattah, and Mamshad Nayeem Rizve. An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized rgb color space. *Journal of healthcare engineering*, 2018, 2018.

[108] Paulo Coelho, Ana Pereira, Marta Salgado, and António Cunha. A deep learning approach for red lesions detection in video capsule endoscopies. In *International Conference Image Analysis and Recognition*, pages 553–561. Springer, 2018.

[109]

[110] Yixuan Yuan, Baopu Li, and Max Q-H Meng. Bleeding frame and region detection in the wireless capsule endoscopy video. *IEEE journal of biomedical and health informatics*, 20(2) :624–630, 2015.

[111] Zahra Amiri, Hamid Hassanpour, and Azeddine Beghdadi. Feature selection for bleeding detection in capsule endoscopy images using genetic algorithm. In *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–4. IEEE, 2019.

[112] Shipra Suman, Fawnizu Azmadi Hussin, Aamir Saeed Malik, Nicolas Walter, Khean Lee Goh, Ida Hilmi, et al. Image enhancement using geometric mean filter and gamma correction for wce images. In *International Conference on Neural Information Processing*, pages 276–283. Springer, 2014.

[113] Shipra Suman, Fawnizu Azmadi B Hussin, Nicolas Walter, Aamir Saeed Malik, Shaiw Hooi Ho, and Khean Lee Goh. Detection and classification of bleeding using statistical color features for wireless capsule endoscopy images. In *2016 International Conference on Signal and Information Processing (IConSIP)*, pages 1–5. IEEE, 2016.

[114] Shipra Suman, Fawnizu Azmadi Hussin, Aamir Saeed Malik, Shiaw Hooi Ho, Ida Hilmi, Alex Hwong-Ruey Leow, and Khean-Lee Goh. Feature selection and classification of ulcerated lesions using statistical analysis for wce images. *Applied Sciences*, 7(10) :1097, 2017.

[115] Konstantin Pogorelov, Shipra Suman, Fawnizu Azmadi Hussin, Aamir Saeed Malik, Olga Ostroukhova, Michael Riegler, Pål Halvorsen, Shiaw Hooi Ho, and Khean-Lee Goh. Bleeding detection in wireless capsule endoscopy videos—color versus texture features. *Journal of applied clinical medical physics*, 20(8) :141–154, 2019.

[116] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2) :189–201, 2009.

[117] E Blotta, A Bouchet, V Ballarin, and J Pastore. Enhancement of medical images in hsi color space. In *Journal of Physics : Conference Series*, volume 332, page 012041. IOP Publishing, 2011.