République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE
Présenté en vue de l'obtention du diplôme
Master en informatique
Spécialité : Réseaux et multimedia

THEME

Classification des cancers basée sur la sélection des gènes des données biopuces

Présenté par :

BOUCHELAL Amel

SELAMA Fateh Mohammed Chaouki

Soutenu publiquement le : 22/06/2024

Devant le jury composé de :

Présidente : Dr. Hafida CHELLAKH

Examinatrice : Dr. Lynda SAIFI

Encadreur : Pr. Djaafar Zouache

2023/2024

الإهداء

«وآخر دعواهم أن الحمد لله رب العالمين »

ما سلكنا البدايات إلا بتيسيره، وما بلغنا النهايات إلا بتوفيقه، وما حققنا الغايات إلا بفضله، فالحمد لله حباً وشكراً وامتناناً. الحمد لله على البدء والختام.

لم تكن الرحلة قصيرة ولم يكن ينبغي لها أن تكون كذلك، لم يكن الحلم قريباً ولم يكن الطريق ممهداً بالسهول، ولكن من قال «أنا لها» نالها. وانا ْنلتهاْ وعانقت اليوم مجداً عظيماً، فعلتها بعد أن كانت مستحيلة. كانت الدروب قاسية، فقدت فيها الكثير، ولكني وصلت والحمد لله.

ولهذا أهدي ثمرة جهدي إلى:

من قال فيهما الله تعالى: ْوقضى ربك ألا تعبدوا إلا إياه وبالوالدين إحساناً

إلى من جعل الجنة تحت أقدامها وسهلت لي الشدائد بدعائها، إلى الإنسانة العظيمة التي تمنت أن ترى هذا اليوم «أمي العزيزة».

إلى من كلل العرق جبينه، ومن علمني أن النجاح لا يأتي إلا بالصبر والإصرار، إلى النور الذي أنار دربي، إلى من بذل الغالي والنفيس واستمددت منه قوتي واعتزازي بنفسي، «والدي العزيز».

إلى من عشت معهن أجمل اللحظات، وكن لي السند، إلى من شاركنني كل بسمة ودمعة، وغمرنني بالحب والتوجيه، إلى أخواتي «داهية و ياسمين».

إلى من تحلت بالأخاء، وتميزت بالوفاء، رفيقتي في المشوار «إيمان».

وفي النهاية، أهدي تحياتي إلى أسرتي جميعاً وكل معارفي، لكل من شجعني بكلمة طيبة وكان لي عوناً وسنداً في هذا الطريق .

أمال بوشلال

الإهداء

بسم الله بادئ البدء، وخاتم الختم، وباعث الهمم، ومانح النعم.

إلى من غرس فيّ بذور الطموح، ورعى نبتة التفوق حتى أثمرت: والدي العزيز، الذي كان سندي ودعمي في كل لحظة، فأهديك هذا العمل البسيط الذي ما هو إلا جزء من جهدك وعطاءك.

وإلى أمي الغالية، منبع الحب والحنان، التي لم تبخل عليّ بالدعاء والدعم المستمر، ولها في القلب منزلة لا تعادلها منزلة.

إلى إخوتي الأحباء، الذين كانوا لي خير رفقاء الدرب، وشركاء الأمل، ومعينين في الشدة، ومحفزين في كل مرحلة.

وإلى خطيبتي العزيزة، التي كانت مصدر إلهام وتشجيع لا ينضب، بوقوفها بجانبي وتحفيزها لي بكل حب وصدق، فأهديك هذه المذكرة عرفاناً بجميلك ودعماً لحبنا الذي يزداد تألقاً يوماً بعد يوم.

والى زميلتي في هذا العمل و التي كانت اكبر سند لي في هذا المشوار طيلة ٥ سنوات لك خالص الشكر و العرفان .

لكم جميعاً أهدي هذا العمل، تقديراً وامتناناً، راجياً من الله أن يوفقني دائماً لأكون عند حسن ظنكم، وأن يكلل جهودنا جميعاً بالنجاح والازدهار.

سلامة فاتح محمد شوقي

Remerciement

Nous exprimons notre profonde gratitude envers le Tout-Puissant pour nous avoir accordé la santé, les ressources, la volonté, le courage et les opportunités nécessaires pour mener à bien cette étude et la réaliser avec succès. Tout d'abord, nous tenons à souligner que ce travail n'aurait pas atteint un tel niveau de qualité sans l'aide précieuse et l'encadrement du Dr. ZOUACHE DJAAFAR. Nous vous exprimons notre sincère reconnaissance pour votre encadrement dévoué tout au long de ce projet. Votre expertise, vos conseils éclairés, votre disponibilité constante, votre patience et votre volonté de partager vos connaissances ont été d'une valeur inestimable et ont grandement contribué à notre réussite. Nous vous sommes sincèrement reconnaissants pour votre implication et votre soutien inconditionnel.

Nous souhaitons également remercier le Dr. Benaouda Nadjib pour son aide précieuse.

Nos remerciements vont aussi à l'ensemble de nos professeurs qui ont joué un rôle essentiel dans notre parcours académique. Votre enseignement passionné, vos encouragements et vos conseils avisés nous ont permis de progresser dans notre discipline et de développer nos compétences.

Enfin, nous exprimons notre profonde gratitude envers nos familles qui nous ont toujours soutenus et ont contribué à notre formation à tous les niveaux d'études.

Résumé

Ce mémoire vise à relever un défi majeur dans le domaine de la recherche sur le cancer, à savoir l'identification des gènes les plus pertinents pour la classification des cancers. Pour ce faire, une approche en trois étapes a été adoptée. Tout d'abord, nous avons appliqué des algorithmes de classification directement sur les ensembles de données biopuces. Ensuite, la qualité des données a été améliorée en appliquant des étapes de prétraitement avant de réappliquer les algorithmes de classification. Enfin, les données prétraitées ont été encore améliorées en sélectionnant les gènes les plus pertinents à l'aide de techniques de sélection basées sur le filtre utilisant l'information mutuelle, avant de réappliquer les mêmes algorithmes de classification. Les résultats de cette étude ont révélé que l'algorithme des machines à vecteurs de support a atteint un taux de classification de 100 % avec la plupart des bases de données utilisées après la sélection des gènes pertinents. L'algorithme des réseaux de neurones a également montré de bonnes performances dans la classification des types de cancer.

Mots-clés : Classification des cancers, Sélection des gènes, Sélection par filtre, Information mutuelle, Données biopuces.

Abstract

This thesis aims to address a major challenge in cancer research, namely the identification of the most relevant genes for cancer classification. To achieve this, a three-step approach was adopted. Firstly, classification algorithms were applied directly to biochip datasets. Subsequently, data quality was improved by applying preprocessing steps before reapplying the classification algorithms. Finally, preprocessed data was further enhanced by selecting the most relevant genes using selection techniques based on mutual information filtering, before reapplying the same classification algorithms. The results of this study revealed that the support vector machine algorithm achieved a classification rate of 100% with most of the databases used after selecting the relevant genes. The neural network algorithm also showed good performance in classifying cancer types.

Keywords : Cancer classification, Gene selection, Filter selection, Mutual information, Biochip data.

ملخص

تهدف هذه الأطروحة إلى معالجة تحدٌ رئيسي في مجال أبحاث السرطان، والمتمثل في تحديد الجينات الأساسية لتصنيف السرطانات. وللقيام بذلك، تم اعتماد منهج مكون من ثلاث مراحل. أولاً، قمنا بتطبيق خوارزميات التصنيف على مجموعات البيانات مباشرة. بعد ذلك، تم تحسين جودة البيانات من خلال تطبيق خطوات المعالجة الأولية قبل تطبيق خوارزميات التصنيف عليها. أخيرًا، تم تحسين البيانات المعالجة مسبقًا من خلال اختيار الجينات الأكثر صلة باستخدام تقنيات الاختيار بواسطة المرشح باستخدام المعلومات المتبادلة، قبل إعادة تطبيق نفس خوارزميات التصنيف عليها مرة أخرى. كشفت نتائج هذه الدراسة أن خوارزمية أجهزة دعم المتجهات تميزت بتحقيق معدل تصنيف ١٠٠ ٪ مع معظم قواعد البيانات التي استخدمناها بعد اختيار الجينات ذات الصلة. كما قدمت خوارزمية الشبكات العصبية أداءً جيدًا في تصنيف أنواع السرطان.

الكلمات المفتاحية: تصنيف السرطان، اختيار الجينات، اختيار بواسطة المرشح، المعلومات المتبادلة، بيانات الرقاقة الحيوية.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Le cancer, en tant que maladie grave caractérisée par une prolifération incontrôlée de cellules pouvant envahir les tissus sains, constitue l'une des principales causes de mortalité à l'échelle mondiale. Sa diversité est reflétée par les nombreux types de cancers classés selon leur origine tissulaire, tels que les cancers du sein, du poumon ou les leucémies, et ses causes multifactorielles sont enracinées dans des facteurs génétiques, environnementaux et liés au mode de vie. La détection précoce revêt une importance capitale, car elle accroît significativement les chances de guérison [1].

Les avancées en biologie moléculaire ont mis en lumière que des altérations génétiques, comme des mutations ou des remaniements chromosomiques, peuvent perturber les mécanismes régulant la croissance et la division cellulaires, menant ainsi à la formation de cellules cancéreuses incontrôlables. L'identification des gènes impliqués dans ces processus est cruciale pour le diagnostic précoce et le développement de traitements ciblés. Cependant, l'abondance de données génomiques issues des technologies de séquençage pose un défi majeur pour la sélection des gènes pertinents distinguant les échantillons cancéreux des sains [2].

Dans cette perspective, la sélection de gènes joue un rôle primordial dans la détection et le diagnostic des cancers. Malgré les dizaines de milliers de gènes dans le génome humain, seule une fraction est réellement pertinente pour caractériser et différencier les tissus cancéreux des tissus sains. Cependant, le bruit et la redondance inhérents aux données génomiques compliquent cette tâche [3].

La problématique centrale de cette étude est donc la suivante : comment identifier efficacement le sous-ensemble de gènes les plus informatifs et pertinents, pour une détection précise et fiable des différents types de cancers ?

Pour répondre à cette problématique, nous avons proposé un processus structuré impliquant l'application et la comparaison d'algorithmes de classification sur des ensembles de données génétiques à travers trois scénarios distincts. Dans un premier temps, les algorithmes de classification (arbre de décision, k-plus proches voisins, réseaux de neurones, machines à vecteurs de support) sont appliqués directement sur les données brutes. Ensuite, les données sont prétraitées (normalisation, gestion des valeurs manquantes, réduction du bruit) avant l'application des mêmes algorithmes. Enfin, les données prétraitées sont optimisées avec des techniques de sélection de gènes (maximisation de l'information mutuelle, maximisation du score de pertinence minimum-redondance, information mutuelle conjointe maximale) avant l'application des algorithmes de classification. L'efficacité de chaque méthode est évaluée en comparant le taux de classification et des métriques de performance complémentaires telles que la précision, le rappel et le score F1. Cette approche systématique vise à améliorer la précision de la détection du cancer et à offrir une évaluation complète des performances des différents algorithmes, fournissant ainsi une méthodologie optimisée pour l'identification des gènes les plus pertinents.

Ce mémoire est structuré en quatre chapitres distincts, chacun explorant un aspect spécifique du travail effectué :

- Le premier chapitre traite des principes, des objectifs et des principales méthodes de la classification supervisée, détaillant le fonctionnement et les spécificités de chaque approche.
- Le deuxième chapitre est consacré au processus de sélection d'attributs, présentant diverses méthodes et soulignant leur importance pour le développement de modèles performants et généralisables.
- Le troisième chapitre explore les aspects biologiques de divers cancers et propose un système de classification en trois étapes pour améliorer le diagnostic à l'aide d'algorithmes de classification.
- le quatrième chapitre détaille les outils, les bases de données génomiques et la méthodologie utilisée pour classifier les données du cancer, en appliquant et comparant des algorithmes selon trois scénarios différents, et analyse les résultats pour identifier les approches les plus performantes.

Chapitre 1

Classification supervisée

1.1 Introduction

La classification supervisée est une technique d'apprentissage automatique qui permet de prédire la classe d'un objet en se basant sur des données d'entraînement préalablement étiquetées. Dans ce chapitre, nous allons explorer les différentes facettes de la classification supervisée. Nous commençons par définir ce qu'est la classification supervisée et son objectif. La classification supervisée consiste à attribuer une classe ou une catégorie à de nouvelles observations, sur la base d'un modèle appris à partir de données d'apprentissage comportant des observations dont la classe est connue. L'objectif est de prédire la valeur d'une variable cible catégorielle en capturant les relations entre cette variable et les attributs descriptifs des observations. Ensuite, nous expliquerons le fonctionnement de cette technique et les différentes méthodes utilisées pour la classification supervisée, notamment l'arbre de décision, les k plus proches voisins, les réseaux de neurones, la méthode de noyau et les SVM. Le fonctionnement général consiste à entraîner un algorithme de classification sur des données étiquetées, afin qu'il puisse apprendre à établir des règles de décision pour prédire les classes des nouvelles données. Différents algorithmes existent, se basant sur des approches statistiques, géométriques, neuronales ou encore sur la théorie des marges. Nous détaillerons le principe et les particularités de chaque méthode : l'arbre de décision qui partitionne récursivement l'espace de données, la méthode des k plus proches voisins qui classe en fonction des points les plus proches, les réseaux de neurones capables d'apprendre des relations non linéaires complexes, les méthodes à noyau

qui projettent les données dans un espace de redescription, et enfin les SVM qui déterminent l'hyperplan optimal pour la séparation des classes.

1.2 Définition de la classification supervisée

La classification supervisée consiste à assigner des objets à des classes prédéfinies, sur la base de leurs caractéristiques descriptives. On définit :[4] :

- Π : la population d'objets.
- D : l'ensemble des descriptions d'objets.
- C : l'ensemble des classes.
- X : la fonction qui associe une description à chaque objet.
- Y : la fonction qui associe une classe à chaque objet.
- F :la fonction de classification recherchée.

L'objectif de la classification supervisée est de trouver la fonction de classification F de telle sorte que $F \circ X$ soit une bonne approximation de Y (voir figure 1.1). En d'autres termes, pour tout $p \in \Pi$, $F(X(p))$ doit être très proche de $Y(p)$, et idéalement, $Y(p) = F(X(p))$. Le processus de classification supervisée comprend une phase d'apprentissage visant à extraire des informations pertinentes à partir de vastes ensembles de données complexes ou à apprendre des comportements à partir d'exemples [4].

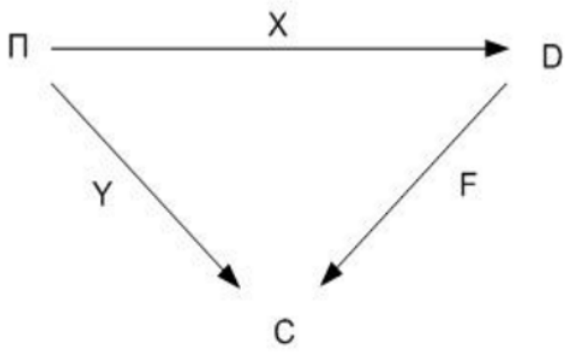

FIGURE 1.1 – fonction de classement.

L'apprentissage, dans ce contexte, correspond à l'acquisition d'une connaissance explicite par un système informatique, impliquant la construction de nouvelles connaissances ou l'amélioration de connaissances existantes. deux types d'apprentissage sont identifiés : l'apprentis-

sage hors ligne, comprenant une période de formation où l'apprenant induit une hypothèse à partir d'un ensemble d'exemples, suivie d'une période d'évaluation où la performance est évaluée avec un autre ensemble d'observations distinct ; et l'apprentissage en ligne, impliquant que l'apprenant reçoit une nouvelle observation, prend une action en conséquence, puis ajuste son hypothèse en fonction de la réponse de l'environnement[4].

1.3 L'objectif de la Classification supervisée

Le premier objectif de la classification supervisée consiste à établir des règles discernantes pour catégoriser des objets en classes distinctes en fonction de leurs caractéristiques, qu'elles soient qualitatives ou quantitatives. ces méthodes peuvent également être étendues pour traiter des variables Y quantitatives, ce qui revient à une forme de régression. initialement, un ensemble de données d'apprentissage est nécessaire, comprenant des exemples dont les classifications sont déjà connues. ce corpus est alors utilisé pour élaborer les règles de classification. l'étape suivante consiste à évaluer la performance et la fiabilité de ces règles, en les comparant et en les appliquant. cette évaluation permet de détecter d'éventuels cas de sous-apprentissage ou de sur-apprentissage, ce dernier étant lié à la complexité du modèle. souvent, un second ensemble de données, indépendant du premier et appelé ensemble de validation ou de test, est utilisé à cette fin.

L'objectif de la classification supervisée est d'apprendre, à l'aide d'un ensemble d'entraînement (ensemble d'apprentissage), une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe [5].

1.4 Fonctionnement de la classification supervisée

Le mécanisme opérationnel de la classification supervisée peut être divisé en deux étapes distinctes :

• La première étape, connue sous le nom de phase d'apprentissage (ou modèle d'apprentissage), englobe tout ce que l'algorithme de classification assimile et représente sous la forme de règles de classification. Ces règles sont essentielles pour établir des associations entre les objets et les classes de référence préalablement identifiées. Pendant cette phase, l'algorithme apprend

à partir d'un ensemble de données d'apprentissage, construisant ainsi le modèle. Deux types de raisonnements caractérisent cette phase : le raisonnement inductif, qui part du particulier pour arriver au général en considérant un ensemble maximal de règles de classification et en le réduisant au mieux, et le raisonnement déductif, qui construit les règles une par une, du général au particulier, jusqu'à obtenir une description précise de l'ensemble d'apprentissage [6].

• La deuxième étape, la phase de classification proprement dite, consiste à utiliser des données de test pour évaluer la précision des règles de classification générées lors de la première phase. Si la précision du modèle est jugée satisfaisante, les règles peuvent être appliquées à de nouvelles données. Le modèle d'apprentissage ainsi élaboré est alors utilisé pour classer de nouveaux objets [6].

1.5 Les méthodes de Classification supervisée

1.5.1 L'arbre de décision

Les arbres de décision sont reconnus pour leur robustesse et leurs performances éprouvées dans diverses applications industrielles et de recherche. Malgré leur histoire conséquente, les algorithmes de construction d'arbres de décision demeurent largement inchangés, visant à segmenter chaque nœud de manière pertinente à l'aide d'un critère spécifique et à construire l'arbre de manière récursive de la racine aux feuilles [7].

Un arbre de décision, dans le domaine de l'informatique, est une méthode de classification en forme d'arbre, composée d'un nœud racine, de nœuds internes représentant les tests sur les fonctionnalités, de branches représentant les résultats des tests, et de feuilles représentant les valeurs prédites. La construction de l'arbre divise récursivement l'espace des fonctionnalités en partitions binaires, en maximisant un critère de pureté des classes à chaque étape [8].

La construction de l'arbre est essentielle dans cette méthode de classification. Les algorithmes descendent lors de la construction de l'arbre, divisant l'échantillon à chaque étape en sous-ensembles les plus homogènes possibles. Cette procédure récursive suit un processus de la racine vers les feuilles, sélectionnant à chaque nœud la division la plus pertinente en fonction de l'attribut-test le plus discriminant [4].

La figure 2 illustre un exemple d'arbre de décision [4].

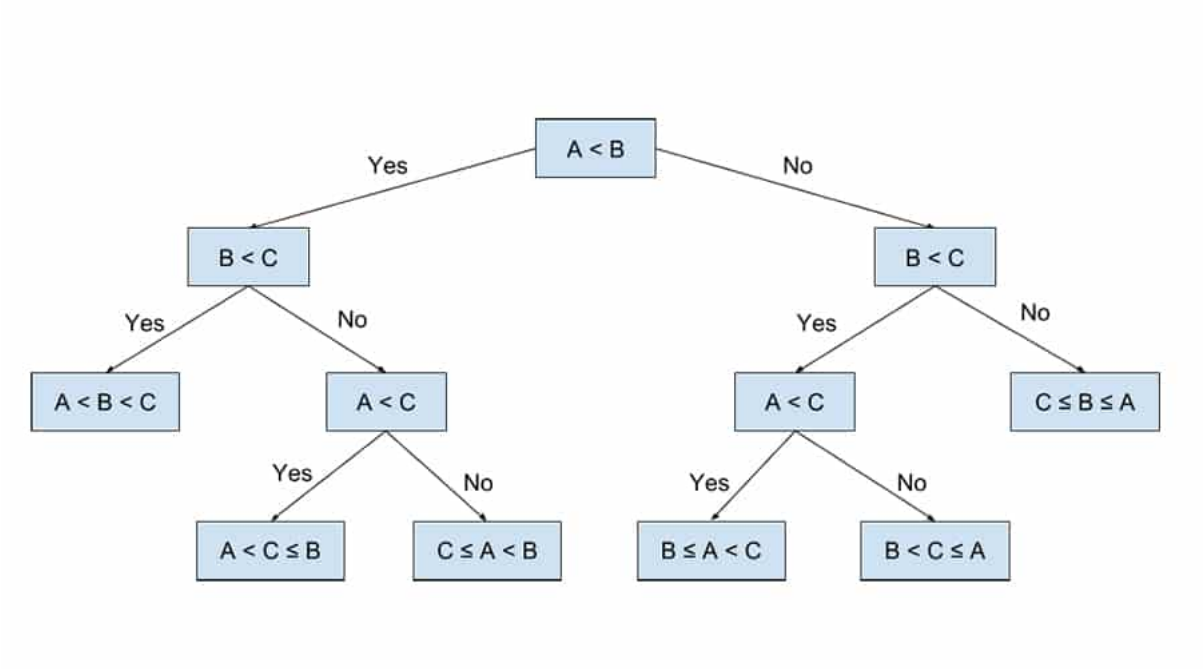

FIGURE 1.2 – Aperçu du fonctionnement d'un arbre de décision.

1.5.1.1 Les avantages de L'arbre de décision

Les arbres de décision sont construits en suivant un ensemble de règles très explicites, ce qui facilite la compréhension des résultats par l'utilisateur. Ils nécessitent généralement peu de ressources, ce qui se traduit par des temps d'apprentissage et de test relativement courts [9].

1.5.1.2 Les inconvénients de L'arbre de décision

Lorsque les arbres de décision deviennent très grands, avec de nombreuses feuilles, leur pouvoir explicatif diminue. De plus, le fait que les sous-nœuds dépendent directement du nœud racine rend le classement très sensible à la présence ou à l'absence d'un seul descripteur dans le texte. Par ailleurs, la modification d'un seul nœud, s'il est proche de la racine, change complètement l'arbre [9].

1.5.2 Classifieur bayésien naïf

Le classifieur naïf bayésien est un modèle d'apprentissage supervisé basé sur le théorème de Bayes. Contrairement à la statistique classique qui estime des paramètres fixes à partir de données, l'approche bayésienne utilise des paramètres aléatoires décrits par des distributions de probabilités initiales (a priori). La formule de Bayes permet de calculer la probabilité finale (a posteriori) en combinant ces probabilités initiales avec les nouvelles données répertoriées.

Ainsi, ce modèle peut être amélioré en intégrant de nouvelles données, ce qui affine les probabilités et l'apprentissage. Une comparaison entre les approches classique et bayésienne est présentée dans la contribution citée [10].

1.5.2.1 les avantages du classifieur bayésien naïf

La simplicité et l'efficacité du classifieur naïf bayésien découlent de l'hypothèse d'indépendance entre les descripteurs. Il se distingue par sa capacité à fonctionner efficacement même avec un petit nombre de documents d'entraînement. Il a démontré son efficacité dans la classification de documents courts, tels que les e-mails, notamment dans la distinction entre les courriels légitimes (ham) et les pourriels (spam) [9].

1.5.2.2 Les inconvénients de Classifieur bayésien naïf

À la différence des documents courts, les documents longs représentent un défi majeur pour le classifieur naïf bayésien, car un vocabulaire étendu favorise les dépendances entre les termes[9].

1.5.3 L'algorithme k plus proches voisins (k-PPV)

L'algorithme des k plus proches voisins (noté k-PPV) est la méthode de classification des plus élémentaires. Elle fonctionne sur la base d'un ensemble d'échantillons appartenant à des classes connues, où un échantillon inconnu est attribué à la classe possédant les k échantillons les plus similaires. Le principe est le suivant : étant donné un nouvel exemple x décrit par p attributs, on identifie dans l'ensemble d'apprentissage les k exemples les plus proches de x. On assigne alors à x la classe majoritaire parmi ces k plus proches voisins. Cela dépend de la méthode de trois paramètres [11] :

- Le nombre de voisins k.
- La mesure de distance entre les exemples, généralement euclidienne sur des attributs numériques.
- La règle de décision : la classe majoritaire parmi les k voisins.

Généralement, on choisit un nombre de voisins entre 1 et 7, et la distance euclidienne est souvent utilisée pour mesurer la proximité entre les exemples lorsque les attributs sont numériques. En ce qui concerne la combinaison des valeurs associées aux voisins pour obtenir la

valeur associée à x, pour la classification, la classe majoritaire parmi les voisins est retenue pour x [11].

L'avantage est la simplicité, pas besoin d'apprendre de modèle. Mais bien que cette méthode soit simple et souvent dotée d'un bon pouvoir prédictif, sa performance diminue avec l'augmentation de la dimension, car il est nécessaire de calculer toutes les distances à chaque nouvelle classification. De plus, son efficacité dépend fortement de k et nécessite un nombre important d'observations pour garantir une précision satisfaisante des résultats [11].

Retenons aussi l'exemple présenté en [12] :

Deux paramètres sont à définir pour appliquer l'algorithme k-PPV :

- k, le nombre de voisins à considérer
- L, la marge d'erreur autorisée

Par exemple, si k=3 et l=3, les 3 plus proches doivent voisins appartenir à la même classe. Si l=2, au moins 2 des 3 plus proches voisins doivent être dans la même classe.

L'algorithme k-PPV peut facilement séparer plusieurs classes et distinguer des classes non linéaires. Son principal inconvénient est sa sensibilité aux valeurs aberrantes parmi les données.

1.5.3.1 Les avantages des K-plus proches voisins (KNN)

- **Simplicité et facilité d'implémentation :** KNN est facile à comprendre et à mettre en œuvre. Il ne nécessite pas de paramètres complexes ou d'hypothèses spécifiques.
- **Adaptabilité aux données changeantes :** KNN peut s'adapter dynamiquement aux nouvelles données sans nécessiter de réapprentissage complet du modèle.
- **Pas de modélisation mathématique :** Contrairement à certains autres algorithmes, KNN n'impose pas de modélisation mathématique préalable [13].

1.5.3.2 Les inconvénients des K-plus proches voisins (KNN)

- **Sensibilité aux valeurs aberrantes :** KNN est sensible aux valeurs aberrantes dans les données. Des points atypiques peuvent influencer les prédictions.
- **Calcul intensif :** Le calcul de la distance entre les points de données peut être coûteux, surtout pour de grandes bases de données.

- **Choix du paramètre K :** La sélection du nombre de voisins (paramètre K) peut être délicate. Une valeur trop petite peut entraîner un surajustement, tandis qu'une valeur trop grande peut entraîner un sous-ajustement [13].

1.5.4 Les réseaux de neurones

Un réseau de neurones est un système de traitement de l'information inspiré par le fonctionnement des neurones biologiques. Il est composé de neurones artificiels interconnectés qui traitent l'information de manière parallèle et distribuée [14].

Le réseau apprend en ajustant les poids des connexions entre les neurones, appelées synapses, qui stockent les connaissances, semblables aux synapses biologiques [14].

Les réseaux neuronaux ont plusieurs caractéristiques intéressantes. Leur non-linéarité provient des neurones, permettant une modélisation complexe adaptée à des problèmes difficiles. L'apprentissage adaptatif se fait en ajustant les poids synaptiques en réponse aux exemples, ce qui améliore le réseau avec le temps. La tolérance aux pannes est une autre qualité, car la connaissance est répartie de manière à rendre le réseau robuste face aux défaillances de certains neurones. Enfin, la capacité de traitement parallèle permet au réseau de traiter plusieurs informations en même temps, augmentant l'efficacité dans le traitement des données [14].

1.5.4.1 Architecture des réseaux de neurones artificiels

Les réseaux de neurones s'organisent en empilant des couches de neurones. On distingue classiquement trois types de couches [15] :

- **La couche d'entrée** reçoit les données à traiter et les transmet aux couches suivantes.
- **Les couches cachées**sont des couches intermédiaires entre l'entrée et la sortie. Le traitement non linéaire de l'information s'effectue au niveau des couches cachées.
- **La couche de sortie**restitue le résultat des traitements sous forme de prédictions ou de décisions.

Chaque couche est composée de plusieurs neurones qui transforment l'information avant de la transmettre à la couche suivante. Ce flux d'information unidirectionnel, de la couche d'entrée vers la couche de sortie en traversant les couches cachées, caractérise l'architecture d'un réseau de neurones. Le nombre de couches cachées et le nombre de neurones par couche déterminent la capacité du réseau à modéliser des problèmes complexes (voir figure 1.3) [15].

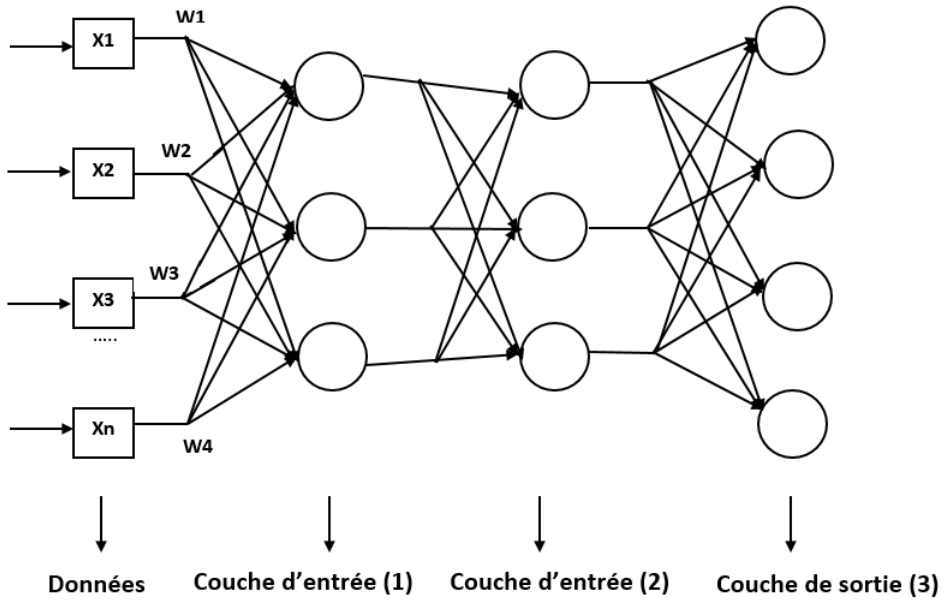

FIGURE 1.3 – Architecture de réseau de neurones.

1.5.4.2 Les avantages des réseaux de neurones

- Capacité à modéliser n'importe quelle fonction, linéaire ou non, simple ou complexe.
- Apprentissage automatique à partir d'exemples représentatifs, par rétropropagation du gradient.
- L'apprentissage est entièrement automatisé.
- Robustesse au bruit et aux données peu fiables.
- Facilité d'utilisation, beaucoup moins de travail manuel que dans l'analyse statistique classique.
- Meilleures performances en cas de peu de données[16].

1.5.4.3 Les inconvénients des réseaux de neurones

- Absence de méthode systématique pour définir la topologie optimale du réseau et le nombre de neurones dans les couches cachées.
- Risques de généralisation et de conclusions précipitées.
- Difficulté d'interprétation du comportement.
- Choix délicat des valeurs initiales des poids synaptiques et du taux d'apprentissage, qui

influencent la vitesse de convergence [16].

1.5.5 Méthodes à noyau et SVM

Les séparateurs à vaste marge ou Support Vector Machines (SVM) sont des classifieurs introduits dans les années 1990, reposant sur deux concepts clés : la notion de marge maximale et l'utilisation de fonctions noyaux. Le principe est de déterminer l'hyperplan séparateur entre les exemples positifs et négatifs qui maximise la marge, c'est-à-dire la distance entre la frontière de séparation et les échantillons les plus proches appelés vecteurs supports. Ceci est formulé comme un problème d'optimisation quadratique. Comme les données ne sont pas toujours linéairement séparables, l'astuce est de projeter les données dans un espace de plus grande dimension, où il devient probable qu'il existe un séparateur linéaire. Pour cela, on utilise une fonction noyau qui réalise cette projection implicitement, sans avoir à calculer les coordonnées des points dans cet espace de grande dimension. Les SVM sont donc des classifieurs performants grâce à l'idée d'espace transformé par noyau permettant la séparation linéaire de données complexes [17].

1.5.5.1 Les avantages des méthodes à noyau et SVM

- Solution optimale globale garantie grâce à la convexité de la fonction objectif.
- Bonnes capacités de généralisation grâce à la maximisation des marges.
- Flexibilité dans le choix des fonctions noyau pour s'adapter à différents types de données.
- Robustesse et résistance au sur-apprentissage grâce aux notions de marge maximale et de vecteurs de support [18].

1.5.5.2 Les inconvénients des méthodes à noyau et SVM

- Temps de calcul important pour l'entraînement, notamment pour de grands jeux de données.
- Difficulté dans le choix des hyperparamètres (fonction noyau, paramètres du noyau, terme de régularisation, etc.).
- Complexité de la fonction de décision qui augmente avec la taille de l'échantillon d'apprentissage.

- Difficulté à appréhender les résultats qui ne fournissent pas de probabilités a posteriori [18].

1.6 Conclusion

La classification supervisée, une technique fondamentale d'apprentissage automatique, se focalise sur la prédiction de la classe d'un objet en utilisant des données d'entraînement préalablement étiquetées. Nous avons mis en lumière son objectif principal qui consiste à développer un modèle capable d'associer de manière précise et fiable les caractéristiques d'entrée aux classes de sortie. Le fonctionnement de cette approche ainsi que les diverses méthodes telles que l'arbre de décision, les k plus proches voisins, les réseaux neuronaux, la méthode de noyau et SVM ont été présentés. Chacune de ces méthodes présente ses propres avantages et limites, leur sélection dépendant du type de données et de la nature du problème à résoudre. Globalement, la classification supervisée se révèle être une approche puissante du machine learning, particulièrement utile pour résoudre une variété de problèmes concrets tels que la détection de défauts d'usinage, de fraudes, de maladies, le tri automatique de courrier, de documents, de vidéos, la reconnaissance d'images, et bien d'autres applications réelles.

Chapitre 2

La selection des attributs

2.1 Introduction

La sélection d'attributs est une étape cruciale de la fouille de données, éliminant les données redondantes pour améliorer les algorithmes de classification. Elle est essentielle pour réduire les temps d'apprentissage et éviter le sur-apprentissage. Cette pratique, utilisée dans divers domaines, nécessite une connaissance approfondie du domaine et peut être effectuée manuellement ou à l'aide d'outils. Deux approches principales, l'encapsulation et le filtrage, sont utilisées. La recherche continue à trouver un sous-ensemble optimal répondant à divers objectifs.

Dans ce chapitre, nous explorerons en détail le processus de sélection d'attributs, en mettant l'accent sur différentes méthodes. Nous aborderons la sélection d'attributs sous différents angles, de la catégorisation des méthodes à la discussion approfondie sur les méthodes de filtrage, de wrapper et intrinsèques (embedded). Enfin, nous examinerons les considérations cruciales lors du choix d'un modèle de sélection d'attributs, soulignant l'importance de cette étape dans le développement de modèles performants et généralisables.

2.2 Définition de la sélection d'attributs

La sélection d'attributs (également appelée sélection de sous-ensembles, sélection de variables ou sélection de caractéristiques) est un processus visant à choisir un sous-ensemble optimal de variables pertinentes parmi un ensemble initial de variables, en se basant sur un

critère de performance spécifique. Il convient de souligner que l'optimalité d'un ensemble de descripteurs n'implique pas nécessairement l'inclusion exclusive des variables considérées comme pertinentes ; il peut également englober des variables non pertinentes qui présentent de meilleures performances en combinaison avec d'autres variables. Ainsi, la sélection de variables vise à identifier le sous-ensemble le plus restreint tout en préservant la précision de la classification et en maintenant la distribution des classes proche de l'originale. Idéalement, les méthodes de sélection d'attributs recherchent le sous-ensemble optimal parmi tous les sous-ensembles candidats, mais cette approche exhaustive peut s'avérer coûteuse. D'autres méthodes adoptent des approches aléatoires et heuristiques pour réduire la complexité, nécessitant un critère d'arrêt afin d'éviter une recherche exhaustive. Trois questions clés émergent à ce stade pour définir les éléments essentiels d'une procédure de sélection de variables [19] :

- Comment mesurer la pertinence des variables ?
- Comment former le sous-ensemble optimal ?
- Quel critère d'optimalité utiliser ?

La pertinence d'une variable est évaluée en fonction de son pouvoir discriminant ou prédictif, exigeant une mesure de pertinence ou un critère d'évaluation. De plus, la procédure de recherche ou de construction du sous-ensemble optimal requiert un critère d'arrêt basé sur la combinaison de la mesure de pertinence et de la procédure de recherche. Une procédure de sélection d'attributs peut être décomposée en quatre étapes types [19] :

- La procédure de génération.
- La fonction d'évaluation.
- Le critère d'arrêt.
- La procédure de validation.

Retenons également la définition présentée dans [20] :

Dans le contexte d'un ensemble de caractéristiques $F = f1, f2, \ldots, fi, \ldots, fn$, l'objectif est de déterminer un sous-ensemble F' qui optimise les performances de l'algorithme d'apprentissage. Formellement,F' doit maximiser une fonction de score υ, de la manière suivante :

F' = argmaxG$\in \Gamma \{\upsilon(G)\}$

Ici, $G \subseteq F$ (avec G « F)peut être un nombre défini par l'utilisateur ou contrôlé par des méthodes de génération de sous-ensembles à explorer ultérieurement. L'espace Γ représente l'ensemble de tous les sous-ensembles possibles de caractéristiques de F [20].

Les techniques de sélection de caractéristiques n'altèrent pas les représentations originales des caractéristiques, mais optent plutôt pour un sous-ensemble parmi celles-ci. Ces approches préservent la sémantique initiale des caractéristiques, offrant ainsi la possibilité d'une interprétation plus aisée par un expert du domaine. En théorie, l'objectif est de trouver le sous-ensemble optimal des caractéristiques qui maximise la fonction de score mentionnée précédemment. Il est crucial que la sélection des caractéristiques soit effectuée uniquement sur les données d'apprentissage, tandis que l'ensemble de test est ultérieurement utilisé pour évaluer la qualité des caractéristiques sélectionnées (sous-ensembles) [20].

2.3 L'objectif de la sélection d'attributs

La sélection d'attributs cherche à identifier un sous-ensemble optimal selon deux critères principaux : il doit contenir des attributs pertinents tout en évitant ceux qui sont redondants. De plus, ce sous-ensemble doit être conçu de façon à optimiser la réalisation de l'objectif fixé. Cet objectif peut être d'atteindre une grande précision d'apprentissage, d'accélérer le processus d'apprentissage ou encore d'assurer une applicabilité efficace du classifieur résultant. L'identification de ce sous-ensemble optimal vise donc à la fois à sélectionner les attributs les plus informatifs, discriminants et complémentaires, et à permettre d'optimiser les performances du système en fonction de l'objectif poursuivi [5].

2.4 Les avantages de la sélection d'attributs

- Amélioration de la classification en éliminant les attributs bruitants, permettant ainsi de créer un sous-ensemble plus petit et plus ciblé.
- Meilleure compréhension des phénomènes étudiés grâce à la focalisation sur des attributs pertinents et significatifs.
- Utilisation de petits sous-ensembles d'attributs pour une meilleure généralisation des données, prévenant ainsi le sur-apprentissage.
- Réduction des temps d'apprentissage et d'exécution une fois les meilleurs attributs identifiés, conduisant à un processus d'apprentissage plus rapide et économique.
- Face à des centaines voire des milliers d'attributs, la sélection vise à éviter la corrélation et la redondance, en privilégiant les attributs pertinents qui fournissent le plus

d'informations pour la classification et en évitant la surcharge d'informations [21].

2.5 Pertinence d'un attribut

On distingue deux types de pertinence pour les attributs : forte et faible. Un attribut est fortement pertinent si ses valeurs varient systématiquement avec les valeurs de classe. Ces attributs sont indispensables et doivent être présents dans le sous-ensemble optimal sélectionné, car leur absence entraîne des erreurs dans la prédiction de la variable cible. À l'inverse, un attribut est faiblement pertinent si son importance n'est pas systématique. Bien que pas toujours importants, ces attributs peuvent devenir nécessaires dans le sous-ensemble optimal dans certaines conditions. La sélection d'attributs consiste donc à identifier les attributs fortement et faiblement pertinents pour ne conserver que ceux apportant une valeur prédictive et informative sur la variable à prédire [22].

2.6 Redondance d'un attribut

La redondance entre attributs trouve généralement sa définition dans la corrélation de leurs valeurs : deux attributs sont considérés comme redondants lorsque leurs valeurs sont parfaitement corrélées. Cependant, cette définition ne s'étend pas directement à un sous-ensemble d'attributs. Pour aborder cette problématique, une approche basée sur le principe de couverture de Markov a été élaborée. Elle consiste à examiner les inter-corrélations entre attributs au sein d'un sous-ensemble. Cette méthode, appliquée avec succès dans des contextes tels que les applications biologiques, permet d'identifier les attributs redondants au sein d'un sous-ensemble, préservant ainsi uniquement les attributs réellement informatifs et non redondants pour la prédiction[22].

2.7 Schéma général de la sélection d'attributs

Les algorithmes de sélection d'attributs suivent généralement quatre étapes [23] :

- Génération d'un sous-ensemble d'attributs.
- Évaluation de la pertinence de ce sous-ensemble.
- Définition d'un critère d'arrêt.

- Validation des résultats.

Le processus peut s'arrêter selon différents critères [23] :

- Lorsqu'un nombre prédéfini d'attributs a été sélectionné.
- Lorsqu'un nombre prédéfini d'itérations a été effectué.
- Lorsque l'ajout ou la suppression d'un attribut n'améliore pas le sous-ensemble.
- Lorsqu'un sous-ensemble optimal selon le critère d'évaluation est obtenu.

Les étapes de génération et d'évaluation de sous-ensembles sont répétées jusqu'à l'obtention du meilleur sous-ensemble d'attributs (FIGURE 2.1) [23].

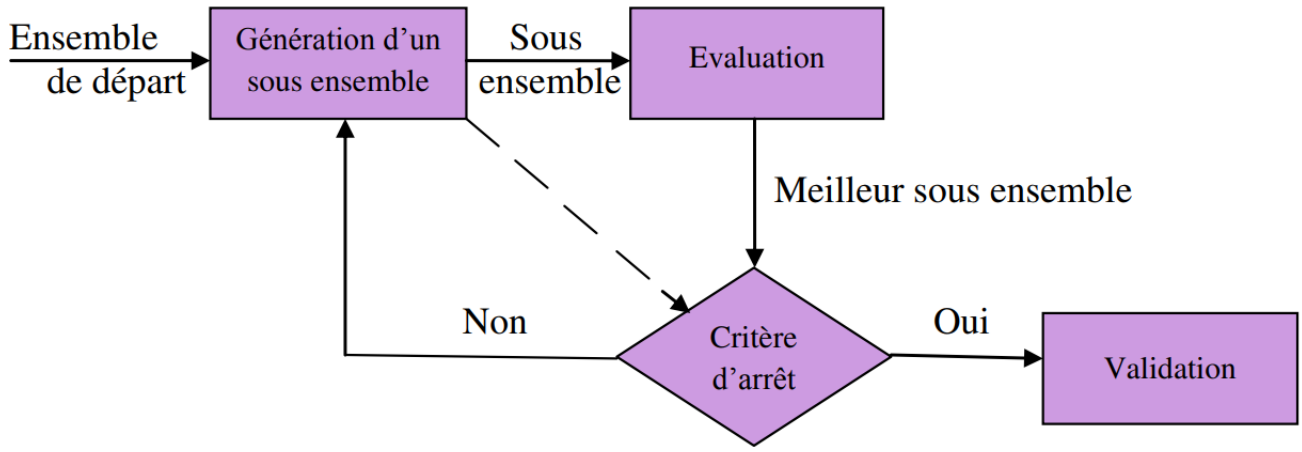

FIGURE 2.1 – Processus de sélection d'attributs avec validation.

2.7.1 Génération de sous-ensemble

Le processus de génération vise à créer un sous-ensemble d'attributs à chaque itération, lequel est ensuite évalué au cours de l'étape d'évaluation. Pour former l'ensemble initial dans cette procédure de génération, plusieurs cas sont étroitement liés à la direction de la recherche suivie ultérieurement [24].

L'ensemble considéré comme point de départ peut être l'ensemble de tous les attributs disponibles, avec les attributs ajoutés itérativement (sélection avant), l'ensemble vide avec les attributs retirés itérativement (sélection arrière), ou un ensemble d'attributs choisis de manière aléatoire, où un nouveau sous-ensemble d'attributs est créé aléatoirement à chaque itération (génération aléatoire) [24].

Une fois l'initialisation effectuée, les stratégies de recherche utilisées pour explorer l'es-

pace des attributs peuvent être classées en trois grandes approches : la génération complète, la génération aléatoire et la génération séquentielle [24].

2.7.1.1 La génération complète

Dans cette méthode de génération complète, on cherche de manière exhaustive le meilleur ensemble de caractéristiques en utilisant différentes façons d'évaluer leur utilité. Même si la recherche peut être complexe, on ne teste qu'un petit nombre d'ensembles possibles. On garantit que l'ensemble de caractéristiques choisi est le meilleur en utilisant des techniques comme le retour en arrière. [25].

2.7.1.2 La génération aléatoire

Dans la recherche aléatoire, il n'y a pas de méthode préétablie pour la sélection. Cette approche implique la création aléatoire de sous-ensembles d'attributs dans le but de choisir le meilleur parmi eux [26].

2.7.1.3 La génération séquentielle

La génération séquentielle implique l'ajout ou la suppression d'un ou plusieurs attributs au fur et à mesure des itérations. Deux approches séquentielles distinctes sont alors distinguées :

- recherche vers l'avant : également appelée approche ascendante, cette méthode débute avec un ensemble vide d'attributs auquel on ajoute un ou plusieurs attributs à chaque itération.
- La recherche vers l'arrière : aussi connue sous le nom d'approche descendante, cette méthode inverse démarre avec l'ensemble complet des attributs. À chaque itération, des attributs sont retirés [22].

2.7.2 Evaluation de sous-ensemble

Chaque ensemble de caractéristiques nouvellement généré doit faire l'objet d'une évaluation selon des critères prédéterminés. La qualité de chaque sous-ensemble est toujours évaluée par rapport à un critère spécifique, ce qui signifie qu'un sous-ensemble choisi peut être optimal selon un critère donné tout en ne l'étant pas en fonction d'un autre. L'évaluation dans les algorithmes de sélection peut être classée en trois catégories distinctes : "filter", "wrapper" et

"embedded" [27] .

2.7.2.1 Les critères de dépendance :

Les critères de dépendance sont fondamentaux dans le cadre de l'algorithme de type Wrapper. L'approche Wrapper repose sur ces critères en utilisant un algorithme de recherche préétabli pour la sélection d'attributs. Cette méthodologie, introduite par John et al. en 1994, implique l'évaluation des sous-ensembles de candidats générés par l'algorithme d'induction. La prédiction de la précision dans la classification peut être employée comme une mesure de dépendance pour la sélection d'attributs. Dans le contexte d'un modèle de sélection d'attributs appliqué à la segmentation, l'objectif est d'évaluer le sous-ensemble optimal d'attributs en fonction de la qualité des clusters obtenus par l'application de l'algorithme de segmentation sur le sous-ensemble sélectionné. Parmi les critères de dépendance les plus renommés, on retrouve les mesures de taux d'erreurs des classifieurs. Les approches Wrappers désignent les algorithmes qui utilisent de tels critères d'évaluation, où le classifieur est considéré comme une fonction d'évaluation [28].

2.7.3 Critères d'arrêt

L'initialisation et le critère d'arrêt jouent un rôle essentiel en délimitant les paramètres de la recherche. Le critère d'arrêt peut être défini de différentes façons [29] :

- Un temps de calcul fixé à l'avance.
- Un nombre d'itérations maximal.
- L'absence d'amélioration de la performance par rapport aux solutions déjà trouvées.
- Le fait que les sous-ensembles candidats deviennent trop homogènes (pour les algorithmes à base de populations).
- Si un sous-ensemble de caractéristiques optimal ou suffisamment bon a été obtenu en fonction des critères d'évaluation.

Ces critères d'arrêt permettent de borner la recherche et d'éviter qu'elle ne se poursuive indéfiniment sans apporter de nouvelles améliorations. L'initialisation et le critère d'arrêt sont donc deux éléments clés pour contrôler l'exploration de l'espace de recherche [29].

2.7.4 Procédures de validation

La validation permet de tester la validité du sous-ensemble d'attributs sélectionnés en effectuant des tests sur des données artificielles ou réelles. Habituellement, l'ensemble de données est divisé en deux sous-ensembles distincts : le sous-ensemble d'apprentissage, composé des prototypes de classes (données avec leurs labels), et le sous-ensemble de test, dont les labels de classes ne sont pas connus. La répartition des données entre ces deux sous-ensembles conduit à différentes approches de validation, notamment [30] :

- **La méthode Holdout :** division des données en deux sous-ensembles, l'un d'apprentissage et l'autre de test, selon des proportions prédéfinies (généralement 70 % pour l'apprentissage et 30 % pour le test).
- **La méthode de resubstitution :**utilisation de l'ensemble d'apprentissage comme ensemble de test.
- **La méthode de V-validation croisée :** partitionnement de l'ensemble des données en V parties quasi-équivalentes, conduisant à V itérations de la procédure de validation où chacune des parties devient l'ensemble de test et les V-1 parties restantes forment l'ensemble d'apprentissage.
- **Le Leave-One-Out :** cas particulier où chaque partie ne contient qu'un exemple.

L'erreur de classification est mesurée sur l'ensemble de test en utilisant les prototypes appris sur l'ensemble d'apprentissage. La validation permet donc d'estimer la capacité de généralisation du modèle [30].

2.8 Les techniques de sélection d'attributs

Les approches de sélection des variables se classifient généralement en trois catégories : les méthodes de filtres, les méthodes d'enveloppes (wrapper) et les méthodes intégrées (embedded). Dans les méthodes intégrées, la sélection des variables s'effectue simultanément à la construction du classifieur par le système d'apprentissage. Cela comporte le risque de surapprentissage, surtout lorsque le nombre d'observations est limité par rapport au nombre de variables. Les méthodes d'enveloppes choisissent un sous-ensemble de variables en évaluant les performances d'un classifieur construit sur ces variables. En raison de l'exploration de l'espace des sous-ensembles de variables, accompagnée de multiples constructions de classifieurs pour une validation croisée, elles impliquent un coût computationnel élevé. Les méthodes de

filtres, en revanche, sélectionnent les variables de manière indépendante du classifieur, ce qui les rend rapides et adaptées au traitement de jeux de données de grande dimension [31].

2.8.1 Les méthodes de filtres :

Dans le cadre des méthodes de filtrage, l'approche adoptée consiste à évaluer chaque attribut de manière univariée afin de lui attribuer un score de pertinence. Ce score facilite le classement des attributs, conduisant ainsi à la sélection des attributs les mieux classés, c'est-à-dire les plus pertinents. Il est pertinent de noter que certaines méthodes multivariées attribuent des scores à des ensembles d'attributs. L'avantage intrinsèque des méthodes de filtrage réside dans leur capacité à être appliquées de manière efficace, même lorsqu'un grand nombre d'attributs est impliqué, en raison de leur complexité raisonnable. Ces méthodes se fondent exclusivement sur les informations disponibles dans les données, demeurant ainsi indépendantes du processus de classification. Cependant, leur principal inconvénient réside dans l'évaluation individuelle des attributs, négligeant les interactions potentielles avec les autres attributs. Par conséquent, ces approches peuvent rencontrer des difficultés lorsqu'il s'agit d'éliminer les attributs redondants. Enfin, ces méthodes reposent généralement sur la pré-détermination d'un seuil pour le critère de pertinence ou d'un nombre d'attributs à sélectionner, rendant le choix de ces paramètres complexe [21].

2.8.2 Les méthodes d'enveloppes (wrapper) :

Les méthodes "wrapper", aussi appelées méthodes enveloppantes, évaluent un groupe de caractéristiques en fonction de leur performance de classification à l'aide d'un algorithme d'apprentissage. La procédure du modèle "wrapper" est illustrée dans la figure 2.2 [32].

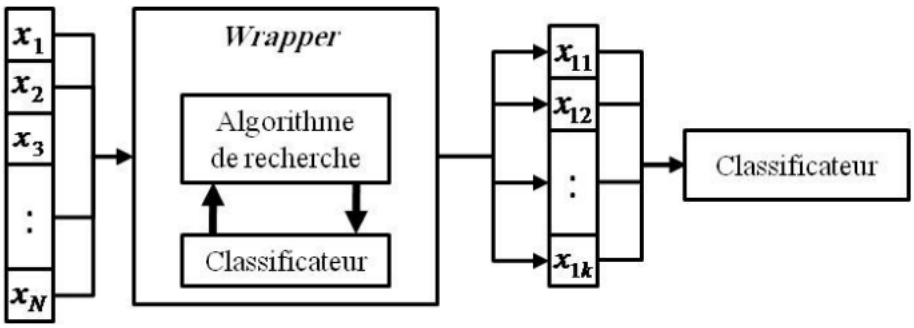

FIGURE 2.2 – La procédure du modèle "wrapper".

On évalue ces caractéristiques avec un classificateur qui estime leur pertinence. Les ensembles de caractéristiques choisis par cette méthode s'adaptent bien à l'algorithme de classification utilisé, mais ils peuvent ne pas être valables si on change de classificateur. Ces méthodes sont coûteuses en temps de calcul en raison de la complexité de l'algorithme d'apprentissage.Pour réduire le temps et éviter le surapprentissage, on utilise souvent la validation croisée. On a montré que les méthodes "wrapper" sont plus performantes que certaines méthodes de filtrage. Cependant, la complexité de cette technique rend impossible l'utilisation d'une stratégie exhaustive en raison de la complexité NP-complète du problème. Ainsi, on peut utiliser des méthodes de recherche heuristiques ou aléatoires. Cependant, à mesure que la taille de l'ensemble initial de caractéristiques augmente, la recherche devient de plus en plus difficile. Les méthodes "wrapper" sont généralement considérées comme meilleures que celles de filtrage. Elles peuvent sélectionner des ensembles de caractéristiques de petite taille performants pour le classificateur utilisé, mais elles ont deux principaux inconvénients : [32] :

a) La complexité et le temps de calcul nécessaires pour la sélection des caractéristiques, beaucoup plus longs que ceux des approches de filtrage et d'autres méthodes de sélection de caractéristiques. [32].

b) La dépendance des caractéristiques sélectionnées au classificateur utilisé. Chaque classificateur ayant ses propres spécificités et hypothèses, le choix du sous-ensemble dépend donc du classificateur choisi. [32].

2.8.3 Les Méthodes Intégrées (Embedded) :

Cette approche combine de manière équilibrée les caractéristiques des méthodes de filtrage et d'enveloppe, offrant ainsi une solution intermédiaire qui résout à la fois la redondance élevée de l'algorithme de filtrage et la complexité de calcul associée à l'algorithme d'enveloppe. La sélection de caractéristiques intégrée se déroule automatiquement pendant le processus d'apprentissage du modèle, ce qui signifie que la recherche et la sélection des sous-ensembles de caractéristiques sont intégrées à la construction du classificateur. Comparées à la méthode d'enveloppe, les approches intégrées présentent des avantages tels qu'une moindre intensité computationnelle, des temps d'exécution plus rapides et un risque potentiellement réduit de surajustement. Toutefois, il convient de souligner que la sélection d'un petit ensemble de caractéristiques peut poser des problèmes dans certains contextes [33].

2.8.4 Les Avantages et les Inconvénients des Méthodes Filtres, Wrapper et Embedded :

Le tableau 2.1 présente les avantages et les inconvénients de ces trois approches [26].

Méthode	Avantages	Inconvénients
Filter	• exécution rapide • coût de calcul faible	• Aucune interaction avec le classificateur
Wrapper	• Interaction avec le classificateur • Bonne performance de classification	• coût de calcul élevé • exécution lente
Embedded	• Interaction avec le classificateur • Bonne performance de classification	• coût de calcul élevé mais plus faible que Wrapper. • exécution lente mais plus rapide que Wrapper. • pas adapté à tous les types de classificateurs.

2.9 Conclusion

La sélection d'attributs constitue une étape cruciale dans le processus de traitement des données et de construction de modèles. Elle vise à identifier les caractéristiques les plus pertinentes tout en minimisant la redondance, contribuant ainsi à la création de modèles plus efficaces et interprétables. En comprenant les objectifs de la sélection d'attributs et en explorant les notions de pertinence et de redondance, nous sommes mieux équipés pour aborder le schéma général de cette procédure. La génération et l'évaluation de sous-ensembles, les critères d'arrêt et les procédures de validation sont autant d'aspects à considérer dans ce processus complexe. Nous avons examiné les diverses techniques de sélection d'attributs, y compris les méthodes de filtres, les enveloppes et les approches intégrées, chaque catégorie offrant ses avantages spécifiques. En synthèse, maîtriser la sélection d'attributs est essentiel pour optimiser la performance des modèles et faciliter leur interprétation dans des domaines variés de l'apprentissage automatique et de l'analyse de données.

Chapitre 3

L'approche proposée pour la détection du cancer

3.1 Introduction

Le cancer représente un défi majeur pour la santé publique mondiale, nécessitant des approches novatrices pour son diagnostic et sa classification. Ce chapitre explore d'abord les aspects biologiques de différents types de cancer tels que le sarcome des tissus mous (SRBCT), la leucémie, le cancer de l'ovaire, le cancer associé au gène MLL, les tumeurs cérébrales, le cancer du poumon, le lymphome diffus à grandes cellules B (DLBCL) et le cancer de la prostate. Ensuite, il propose un système de classification du cancer composé de trois étapes. La première étape consiste en une classification sans prétraitement des données ni application d'algorithmes de sélection de variables. La deuxième étape implique un prétraitement des données mais sans sélection de caractéristiques. Enfin, la troisième étape combine le prétraitement des données avec la sélection de variables pertinentes. Les techniques de prétraitement des données, telles que le nettoyage, la normalisation et la sélection de sous-ensembles de caractéristiques, sont discutées en détail. Les concepts fondamentaux de l'entropie de Shannon et de l'information mutuelle, ainsi que leurs propriétés, sont présentés, conduisant à des techniques de sélection de caractéristiques basées sur des mesures d'information mutuelle comme la réduction de la redondance maximale pertinente (mRMR) et la maximisation de l'information mutuelle (MIM). Enfin, les algorithmes de classification couramment utilisés, tels que les k-plus proches voi-

sins, les machines à vecteurs de support, les arbres de décision et les réseaux de neurones, sont passés en revue et prêts à être appliqués aux données prétraitées.

3.2 L'Aspect Biologique

Le cancer (terme médical : néoplasme malin) est un terme générique pour de nombreuses maladies caractérisées par la croissance incontrôlée d'un groupe de cellules, l'envahissement des tissus adjacents et leur destruction, ou la dissémination dans d'autres parties du corps par les voies lymphatiques ou sanguines. Ces trois caractéristiques définissent le caractère malin des cancers par opposition aux tumeurs bénignes qui n'ont pas ces trois caractéristiques [34].

Voici les définitions de certains types de cancer :

- **SRBCT :**Les petites tumeurs à cellules rondes bleues constituent un groupe de néoplasmes malins hétérogènes caractérisés par des cellules tumorales indifférenciées avec des noyaux petits et hyperchromatiques et un cytoplasme réduit. Ces tumeurs, incluant des types comme les tumeurs d'Ewing, le rhabdomyosarcome et le neuroblastome, présentent souvent des caractéristiques microscopiques difficiles à distinguer, rendant le diagnostic précoce et précis crucial. Pour un diagnostic définitif, des techniques comme l'immunohistochimie et la cytogénétique sont nécessaires [35].
- **Leukemia :** La leucémie est une maladie maligne du sang. C'est la prolifération clonale des cellules souches de la moelle osseuse. Les 4 sous-types les plus fréquemment vus par les omnipraticiens sont la leucémie lymphoblastique aiguë, la leucémie myéloïde aiguë, la leucémie lymphoïde chronique et la leucémie myéloïde chronique. La leucémie lymphoblastique aiguë est plus fréquente chez l'enfant et les autres sont plus fréquentes chez l'adulte [36].
- **Ovarian :**Le cancer de l'ovaire suscite une grande inquiétude en tant que type de cancer gynécologique majeur. Des études ont révélé que le cancer épithélial de l'ovaire n'est pas une entité unique, mais plutôt un ensemble diversifié de tumeurs classifiables en fonction de leurs caractéristiques morphologiques et génétiques distinctes. Les carcinomes séreux de bas grade, les endométrioïdes de bas grade, les carcinomes à cellules claires, les carcinomes mucineux et les tumeurs transitionnelles (de Brenner) font partie d'une catégorie appelée type 1[37] .
- **MLL :**La leucémie lymphoblastique aiguë associée au réarrangement du gène MLL

(LLA-MLL) est une variante agressive de la leucémie lymphoblastique aiguë (LLA) qui montre généralement une réponse relativement faible au traitement, et qui est principalement observée chez les nourrissons (moins de 1 an). Environ 50 % de ces patients nécessitent l'identification de nouvelles cibles thérapeutiques pour améliorer le traitement. Ainsi, il est crucial de comprendre cette maladie et d'identifier les caractéristiques biologiques et moléculaires associées au réarrangement du gène MLL.[38].

- **Prostate-Tumor :**Le cancer de la prostate est une forme de tumeur qui dépend des hormones et provient de l'épithélium, résultant d'une prolifération incontrôlée de cellules instables génétiquement modifiées. Les cellules souches sont un objectif thérapeutique pour le cancer de la prostate. Cependant, puisque le développement de la maladie se déroule sur plusieurs décennies, il est crucial d'identifier et de viser la cellule responsable du renouvellement du cancer (CRC) qui maintient les clones malveillants [39].
- **Lung Cancer :** Le cancer du poumon est une pathologie caractérisée par la prolifération anarchique de cellules pulmonaires anormales qui forment une tumeur. Dans les stades précoces, cette maladie reste souvent asymptomatique. Elle touche principalement les personnes âgées de 60 à 70 ans environ. Le tabagisme chronique représente le principal facteur de risque, même si des cas de cancer du poumon peuvent également survenir chez des individus n'ayant jamais fumé. Une détection précoce est cruciale, les premiers symptômes étant souvent discrets ou absents, permettant ainsi au cancer de progresser silencieusement [40].
- **Diffuse Large B-cell Lymphoma (DLBCL) :**Se caractérise par une prolifération diffuse de grandes cellules lymphoïdes B néoplasiques, dont les noyaux sont de taille égale ou supérieure à celle des noyaux des macrophages normaux, ou mesurant plus de deux fois la taille d'un lymphocyte normal [41].
- **Brain Tumors :**Une tumeur cérébrale est une accumulation de cellules anormales dans le cerveau. Les tumeurs cérébrales peuvent être malignes ou bénignes. Cependant, l'espace dans le crâne est limité. Par conséquent, toute tumeur cérébrale, même bénigne, peut interférer avec les fonctions de votre cerveau et de votre corps. Les tumeurs cérébrales peuvent détruire les cellules cérébrales, augmenter l'inflammation et élever la pression dans le cerveau. Une tumeur cérébrale primaire commence dans votre cerveau. Lorsque des cellules cancéreuses provenant d'autres parties de votre corps causent une tumeur dans votre cerveau, on parle de tumeur cérébrale "secondaire" ou "métastatique".

Les tumeurs cérébrales secondaires sont trois fois plus courantes que les tumeurs cérébrales primaires. Toutes les tumeurs cérébrales secondaires sont malignes. Les tumeurs cérébrales sont décrites par leur localisation, leur type de tissu et les cellules qui composent la masse [42].

3.3 L'approche proposée pour la classification du cancer :

L'approche proposée pour la classification du cancer repose sur une séquence d'étapes. Tout d'abord, un ensemble de données contenant les gènes classés par types de cancer est rassemblé. Ces données subissent ensuite un prétraitement pour les préparer à l'analyse. À l'étape suivante, des algorithmes de classification sont appliqués aux données sans utiliser d'algorithmes de détection. Ensuite, le processus est répété, mais cette fois en appliquant les algorithmes de détection avant l'étape de classification. Cette double approche vise à améliorer la précision de la classification des types de cancer en tirant parti à la fois des algorithmes de classification et de détection de manière intégrée.

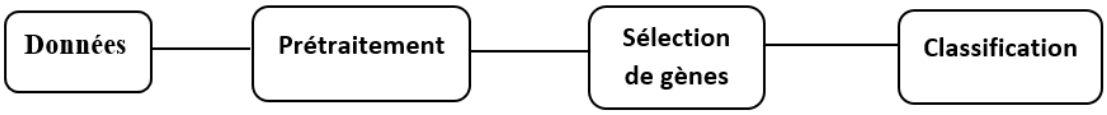

FIGURE 3.1 – L'approche proposée pour la classification du cancer (schéma).

3.4 Prétraitement des données :

La phase de préparation et de nettoyage des données vise à mettre les données d'entrée dans un format adéquat pour l'entraînement des modèles d'apprentissage automatique. En effet, des données brutes, mal formatées ou comportant des erreurs peuvent grandement nuire aux performances du modèle et fausser ses prédictions. Au contraire, un prétraitement rigoureux des données permet d'optimiser l'apprentissage du modèle à partir de ces données et d'obtenir des résultats plus précis et fiables en sortie.

Les étapes que nous avons suivies dans le prétraitement des données sont les suivantes :

3.4.1 Nettoyage des données :

Dans cette partie, nous avons remplacé les valeurs manquantes en utilisant la méthode d'imputation 'SimpleImputer' avec la moyenne de chaque colonne. Cela garantit que toutes les données sont complètes et prêtes à être utilisées pour l'entraînement des modèles de classification.

3.4.2 Normalisation :

La normalisation des données est essentielle pour de nombreux algorithmes d'apprentissage automatique, en particulier ceux basés sur des calculs de distances ou de produits scalaires, comme les SVM (Support Vector Machines) ou les k-plus proches voisins. Elle permet d'éviter que les caractéristiques à grande échelle ne dominent le processus d'apprentissage et améliore les performances du modèle.

3.4.3 Sélection d'un Sous-ensemble de Données :

Nous devions restreindre notre sélection à un sous-ensemble limité de nos données afin de cibler les caractéristiques les plus pertinentes et informatives parmi les données génétiques, et ainsi élaborer des modèles de classification précis et fiables. En travaillant avec des données génomiques ou d'expression génique, la présence d'un grand nombre de caractéristiques (gènes) peut rendre l'entraînement des modèles de classification inefficace et coûteux en termes de calcul, avec un risque accru de surapprentissage. C'est pourquoi l'application de techniques de sélection de caractéristiques est cruciale pour réduire la complexité des données. L'objectif est de ne conserver qu'un sous-ensemble des gènes les plus informatifs et discriminants pour la classification des types de cancer. Ces méthodes permettent d'identifier les gènes les plus significatifs pour distinguer les différentes classes de cancer, tout en éliminant les gènes redondants ou non pertinents. Par conséquent, les performances des modèles sont considérablement améliorées en termes de précision, de rappel et de généralisation.

Dans ce travail, nous utilisons l'information mutuelle comme mesure de pertinence des attributs. Pour cette raison, nous donnons dans cette section un bref aperçu des concepts de base de la théorie de l'information qui sont nécessaires pour comprendre la méthode de sélection des attributs utilisée dans ce travail, à savoir l'information mutuelle.

3.4.3.1 L'Entropie de Shannon :

L'entropie de l'information, également connue sous le nom d'entropie de Shannon en hommage à Claude E. Shannon qui a développé de nombreux concepts fondamentaux de la théorie de l'information, est une mesure de la quantité d'information contenue dans un signal ou un événement. C'est Shannon qui a introduit la notion d'entropie de l'information dans son célèbre article de 1948 intitulé "Une Théorie Mathématique de la Communication". L'idée intuitive derrière l'entropie de l'information est qu'elle représente le degré d'incertitude ou d'imprévisibilité associé à un événement régi par une distribution de probabilités données. Plus la distribution est imprévisible, plus l'entropie est élevée, et vice versa [43].

Pour une variable aléatoire discrète X, l'entropie de X, notée H(X), est définie comme suit[44] :

$$H(X) = -\sum_{x \in X} p(X = x) \log p(X = x)$$

Nous généralisons l'entropie d'une variable aléatoire à une paire de variables aléatoires, en considérant (X, Y) comme une seule variable aléatoire avec des valeurs vectorielles. L'entropie jointe H(X, Y) pour deux variables aléatoires discrètes (X, Y) avec une distribution jointe p(x, y) est définie comme suit [45] :

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

Nous introduisons l'entropie conditionnelle d'une variable aléatoire étant donnée une autre comme la moyenne pondérée des entropies des distributions conditionnelles.

Si (X, Y) suit la distribution conjointe p(x, y), alors l'entropie conditionnelle H(Y|X) est définie comme suit [45] :

$$H(Y|X) = \sum_{x \in X} H(Y|X = x)$$

3.4.3.2 Information Mutuelle :

L'information mutuelle est un indicateur fiable de la pertinence entre les variables, ce qui en fait une mesure largement utilisée dans divers algorithmes de sélection de caractéristiques. Cependant, son calcul peut s'avérer complexe, et l'efficacité d'un algorithme de sélection de caractéristiques est étroitement liée à la précision de l'information mutuelle calculée [40].

L'information mutuelle I(X : Y) quantifie la moyenne de l'information sur la variable X qu'on peut obtenir en connaissant la variable Y [46] :

$$I(X:Y) = \sum_{x,y \varepsilon X} P(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

L'information mutuelle IM est nulle lorsque les variables X et Y sont statistiquement indépendantes, ce qui signifie [47] :

$$p(x_i, y_j) = p(x_i)\,p(y_j)$$

L'information mutuelle est liée de manière linéaire aux entropies des variables selon les équations suivantes [47] :

$$IM(X;Y) = \begin{cases} H(X) - H(X|Y) \\ H(Y) - H(Y|X) \\ H(X) + H(Y) - H(X,Y) \end{cases}$$

Le diagramme de Venn présenté à la Figure 4.3 illustre graphiquement les relations entre l'information mutuelle, l'entropie de X et l'entropie de Y, telles que décrites dans les équations [47].

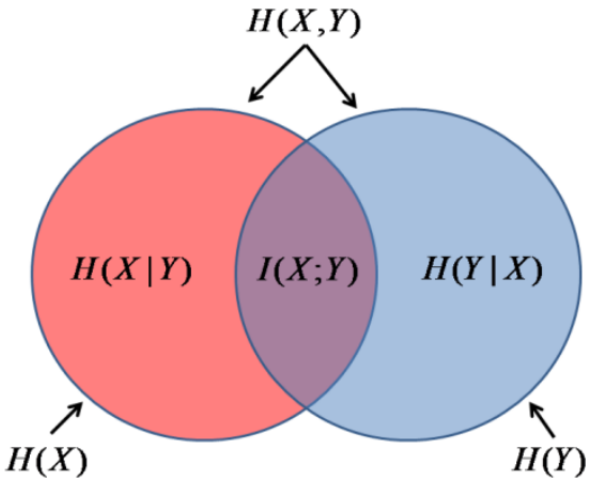

FIGURE 3.2 – Diagramme de Venn : relations entre l'information mutuelle et l'entropie.

Si Z est une variable aléatoire discrète, l'interaction entre Z et les deux autres variables peut

être évaluée en utilisant l'information mutuelle conditionnelle, définie comme suit [47] :

$$IM(X;Y|Z) = H(X|Z) + H(X|Y,Z)$$

La mesure de l'information entre deux variables en présence d'une troisième est possible grâce à l'information mutuelle conditionnelle, mais elle ne capture pas directement l'information partagée entre les trois variables. Pour aborder cette situation, McGill [97] a introduit l'information mutuelle multivariée, une extension significative de l'information mutuelle, permettant d'évaluer l'interaction entre plus de deux variables. En considérant trois variables aléatoires [47].

sa définition est la suivante [48] :

$$IM = (X;Y;Z) = IM(X;Y) - IM(X;Y|Z)$$

$$= H(X) + H(Y) + H(Z) - H(X,Y) - H(X,Z) - H(Y,Z) + H(X,Y,Z)$$

3.4.3.3 Propriétés de l'information mutuelle :

- Pour deux variables aléatoires X et Y, l'information mutuelle $IM(X;Y)$ est toujours positive ou nulle : $IM(X;Y) \geq 0$. Elle est strictement positive sauf dans le cas particulier où X et Y sont statistiquement indépendantes, auquel cas $IM(X;Y) = 0$.
- Lorsqu'on considère trois variables aléatoires X, Y et Z, l'information mutuelle conditionnelle $I(X;Y|Z)$ respecte également $IM(X;Y|Z) \geq 0$. L'égalité $I(X;Y|Z) = 0$ est vérifiée si et seulement si X et Y sont indépendantes conditionnellement à la connaissance de Z [48].

3.5 Sélection d'attributs basée sur le filtre information mutuelle :

L'information mutuelle permet de quantifier la relation de dépendance entre deux variables aléatoires. Dans le cadre de la sélection d'attributs, on évalue l'information mutuelle entre la variable à prédire (variable cible) et chaque variable explicative (attribut) afin de mesurer

leur degré de dépendance mutuelle. Les attributs présentant une forte information mutuelle avec la variable cible sont considérés comme les plus informatifs et pertinents, et seront donc sélectionnés pour être intégrés dans le modèle prédictif.

Plusieurs critères de sélection de caractéristiques basés sur l'information mutuelle ont été proposés dans la littérature scientifique. Pour cette étude, nous avons retenu trois critères particuliers, qui seront détaillés dans les sections suivantes [49] :

- **MRMR (Max-Relevance Min-Redundancy) :** L'algorithme minimal-redundancy-maximal-relevance (mRMR) représente un exemple classique d'algorithme de sélection de caractéristiques. Pour choisir la caractéristique présentant le moins de redondance avec les caractéristiques déjà sélectionnées et la plus grande pertinence avec l'étiquette de classe, la fonction objectif de mRMR soustrait la valeur moyenne de l'information mutuelle entre les caractéristiques de celle entre les caractéristiques et l'étiquette de classe, puis sélectionne la caractéristique avec la plus grande différence [49].
- **JMI (Joint Mutual Information) :** La méthode JMI est une méthode de sélection de caractéristiques qui compare les caractéristiques en fonction de la dominance de l'information mutuelle. Elle utilise une stratégie de recherche gloutonne progressive pour sélectionner les caractéristiques ayant le compte de dominance maximum[50].
- **MIM (Mutual Information Maximization) :** La stratégie de maximisation de l'information mutuelle peut être subdivisée en deux objectifs distincts : maximiser H(Y) et minimiser H(Y|X).Le premier objectif vise à éviter que le modèle ne favorise excessivement une classe spécifique, garantissant ainsi une distribution équilibrée des prédictions. Le second objectif renforce la confiance du modèle dans ses prédictions tout en élargissant la marge entre les classes. Une marge accrue permet à la frontière de décision de mieux refléter la structure réelle du domaine source, favorisant ainsi la performance optimale du même classificateur à travers différents domaines. Cette approche renforce la robustesse du modèle lors des phases de test [51].

3.6 Classification :

Une fois les attributs extraits et la sélection d'attributs basée sur le filtre d'information mutuelle appliquée, la dernière étape consiste à réaliser la classification. En apprentissage automatique, la classification comprend deux phases : l'entraînement et le test. Lors de la phase

d'entraînement, un modèle prédictif est construit à partir des données d'apprentissage. Lors de la phase de test, ce modèle est évalué afin de déterminer s'il est suffisamment précis pour être déployé sur de nouvelles données [4].

Dans notre travail, nous avons opté pour des algorithmes d'apprentissage supervisé pour la tâche de classification, à savoir les k-plus proches voisins, les machines à vecteurs de support, les arbres de décision et les réseaux de neurones. Le fonctionnement de ces algorithmes a été expliqué en détail dans le chapitre précédent.

Dans le chapitre suivant, nous présentons les résultats de notre étude expérimentale sur une base de données génomiques des cancers, en utilisant différentes techniques de sélection d'attributs et différentes méthodes de classification. Les performances des différents modèles seront comparées et analysées afin de déterminer les meilleures pratiques pour cette tâche de classification spécifique.

3.7 Conclusion

Ce chapitre a offert une exploration détaillée des défis inhérents à la classification du cancer, en s'appuyant sur une analyse biologique de divers types de cancer tels que le SRBCT, la leucémie... etc. Il a introduit un système sophistiqué de classification du cancer qui intègre des étapes cruciales de prétraitement, de sélection et de classification des données. À travers le nettoyage, la normalisation et la sélection ciblée des caractéristiques, ce système vise à optimiser la précision diagnostique. La discussion sur l'entropie de Shannon et l'information mutuelle a éclairé leur importance pour la sélection des attributs via des approches telles que Max-Relevance Min-Redundancy, Joint Mutual Information et la Maximisation de l'Information Mutuelle. Enfin, en passant en revue des algorithmes de classification tels que les k-plus proches voisins, les machines à vecteurs de support, les arbres de décision et les réseaux de neurones, ce chapitre pose les bases pour une application efficace de ces méthodologies à des données rigoureusement prétraitées, illustrant une avancée significative dans la lutte contre le cancer.

Chapitre 4

Implementetion et Resultats

4.1 Introduction

Ce chapitre est consacré aux expérimentations réalisées dans le cadre de notre étude sur la classification des données génomiques du cancer. Nous commençons par présenter les outils utilisés pour le développement du projet, y compris le choix du langage de programmation et la configuration matérielle spécifique. Ensuite, une description détaillée des bases de données génomiques du cancer exploitées est fournie, incluant le nombre d'échantillons, d'attributs et de classes. Les concepts d'évaluation des modèles de classification sont abordés, notamment la matrice de confusion et les critères de performance tels que la précision, le rappel et le F1-score, avec l'utilisation de chacun d'eux pour comparer les résultats obtenus dans notre étude. La méthodologie adoptée consiste à appliquer les algorithmes de sélection et de classification sur chaque base de données selon trois scénarios distincts : sans prétraitement ni sélection d'attributs, avec prétraitement seulement, et avec à la fois prétraitement et sélection d'attributs. Les résultats obtenus sont ensuite analysés et comparés afin d'identifier les meilleurs algorithmes pour la classification des données génomiques du cancer.

4.2 Outils matériels et logiciels

4.2.1 Configuration matérielle :

Ce travail a été implémenté sur un PC, caractérisé comme suit :

- **Processeur :** Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz.
- **RAM :** : 8 Go.
- **Système d'exploitation :**64 bits.

4.2.2 Environnement de développement et outils

4.2.2.1 PyCharm :

Dans le cadre du développement de notre application, nous avons opté pour PyCharm en tant qu'environnement de développement intégré (IDE). Créé par JetBrains, cet IDE offre de nombreuses fonctionnalités visant à optimiser le processus de développement. Grâce à PyCharm, nous bénéficions d'un gain de temps considérable lors de la programmation. L'une de ses forces réside dans sa capacité à s'intégrer de manière transparente avec d'autres environnements et outils, facilitant ainsi la collaboration et l'exploitation de différentes ressources. De plus, PyCharm dispose d'un système intelligent de détection et d'installation automatique des packages requis, évitant ainsi toute tâche manuelle fastidieuse.

4.2.2.2 Python :

Python est l'un des langages de programmation les plus prisés et largement utilisés, ce qui s'explique par sa polyvalence lui permettant de s'adapter à divers paradigmes de programmation. Sa simplicité d'utilisation et la richesse de ses bibliothèques, telles que NumPy et Pandas, en font une plateforme de développement fluide et productive. Pour notre projet, nous avons naturellement choisi ce langage, exploitant ses avantages pour répondre efficacement à nos besoins spécifiques.

4.2.2.3 Différentes bibliothèques utilisées

Voici une description des bibliothèques utilisées dans notre code :

- **Pandas :** Pandas est une bibliothèque Python spécialisée et open source pour l'analyse des données. Créée par Wes McKinney en 2008 et ensuite développée avec l'aide de Sien Chang, Pandas est devenue l'une des bibliothèques les plus populaires au sein de la communauté Python. Elle offre une approche simple et complète pour le traitement, l'extraction et la manipulation des données, répondant ainsi à un besoin crucial dans le domaine de l'analyse de données. Fondée sur NumPy, cette bibliothèque bénéficie

non seulement de la compatibilité avec d'autres modules, mais aussi des performances élevées de calcul de NumPy, ce qui a grandement contribué à son succès et à sa diffusion rapide [52].

- **Sklearn (scikit-learn) :** Scikit-learn représente une bibliothèque Python open source dédiée à l'apprentissage automatique. Elle propose une gamme de fonctionnalités pour estimer des modèles tels que les forêts aléatoires, les régressions logistiques, les algorithmes de classification et les machines à vecteurs de support. Cette bibliothèque est conçue pour s'intégrer harmonieusement avec d'autres bibliothèques Python open source, notamment NumPy [53].
- **NumPy :** NumPy représente le fondement du calcul scientifique avec Python, en particulier dans le domaine de l'analyse de données. Cette bibliothèque incontournable sert de socle à de nombreux autres packages Python dédiés aux mathématiques et aux sciences. Parmi ces derniers figure Pandas, une bibliothèque spécifiquement conçue pour l'analyse de données et qui exploite pleinement les concepts introduits par NumPy. En réalité, les outils natifs de la bibliothèque standard Python se révèlent souvent trop limités ou inappropriés pour effectuer la majorité des calculs complexes requis en analyse de données. C'est pourquoi NumPy et ses dérivés comme Pandas sont devenus des incontournables dans ce domaine [54].

4.3 Bases de données utilisées :

Dans le Tableau 4.1, nous présentons les différentes bases de données génomiques de cancer utilisées dans notre étude. Ces bases de données couvrent divers types de cancers, à savoir la leucémie, le cancer du poumon, la leucémie lymphoblastique impliquant le gène MLL, le cancer de l'ovaire, le sarcome des tissus mous, la tumeur cérébrale, le lymphome diffus à grandes cellules B, le cancer du poumon et la tumeur de la prostate. Pour chacune de ces bases de données, nous fournissons des informations clés comprenant le nombre d'échantillons, le nombre d'attributs (gènes) et le nombre de classes.

TABLE 4.1 – Informations sur différents jeux de données.

Nom du dataset	Nombre d'échantillons	Nombre d'attributs	Nombre de classes
Leukemia	72	7129	2
Lung	203	12600	5
MLL	72	12582	3
Ovarian	253	15154	2
SRBCT	83	2308	4
Brain-Tumor1	90	5920	5
DLBCL	77	5469	2
Lung-Cancer	203	12600	5
Prostate-Tumor	103	10509	2

4.4 Évaluation d'un modèle :

L'évaluation des performances d'un modèle prédictif de classification en apprentissage automatique revêt une importance cruciale dans le processus de sélection du modèle optimal. Cette étape permet de juger de la qualité et de la fiabilité des prédictions effectuées par le modèle entraîné. Pour mener à bien cette phase d'évaluation de manière rigoureuse, un outil clé est utilisé : la matrice de confusion. Celle-ci fournit une représentation visuelle détaillée qui confronte les prédictions du modèle aux véritables classes des observations. L'analyse approfondie de cette matrice permet alors de calculer et d'interpréter diverses mesures de performance telles que la précision, le rappel, le score F1 ou encore le taux d'erreurs. C'est sur la base de ces indicateurs quantitatifs que l'expert pourra comparer objectivement différents modèles candidats et retenir celui démontrant les meilleures capacités prédictives sur les nouvelles données [55].

4.4.1 Matrice de Confusion :

Une matrice de confusion est un outil standard pour évaluer les performances d'un modèle de classification. Elle permet de visualiser les prédictions correctes et incorrectes du modèle en comparant les valeurs réelles aux valeurs prédites.

Dans le cas d'un problème de classification binaire, la matrice de confusion est composée de 4 éléments clés :

- **Vrais Positifs (VP) :** Le nombre d'observations positives correctement classées comme positives.
- **Faux Négatifs (FN) :** Le nombre d'observations positives incorrectement classées comme

négatives.

- **Faux Positifs (FP) :** Le nombre d'observations négatives incorrectement classées comme positives.
- **Vrais Négatifs (VN) :** Le nombre d'observations négatives correctement classées comme négatives [56].

En cas de classification multiclasse, les métriques conçues pour la classification binaire ne sont pas directement applicables. Contrairement à la matrice de confusion binaire, la matrice de confusion multiclasse (illustrée dans la Figure 1b) a une dimension N × N, où N représente le nombre d'étiquettes de classe différentes (C0, C1, ..., CN). Ainsi, les concepts de Vrais Positifs (VP), de Vrais Négatifs (VN), de Faux Positifs (FP) et de Faux Négatifs (FN) ne s'appliquent pas de manière directe. Cependant, une analyse spécifique à chaque classe peut être réalisée en se basant sur la structure de la matrice de confusion multiclasse, comme présenté dans la Figure 4.1. En définissant un ensemble de métriques pour chaque classe selon cette approche, il devient alors possible de combiner ces métriques de manière appropriée pour obtenir des mesures globales de performance à partir de la matrice de confusion [57].

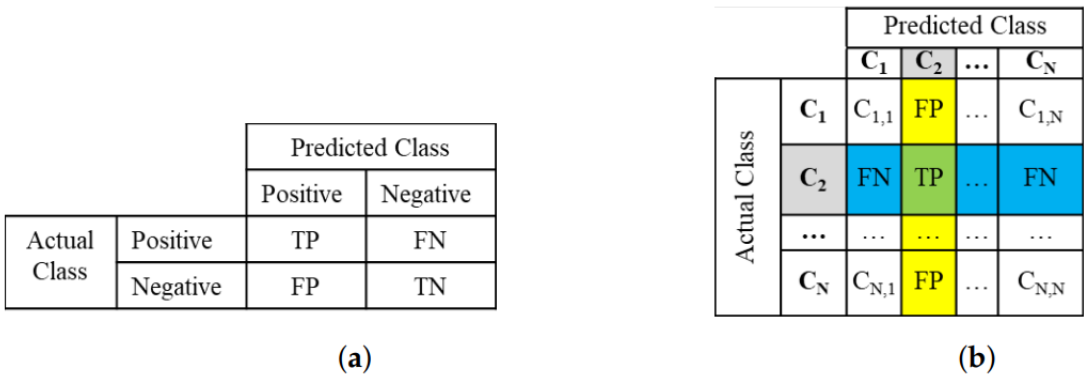

FIGURE 4.1 – Confusion matrix examples. (a) Binary classification problem confusion matrix. (b) Multiclass classification problem confusion matrix.

4.4.2 Critères d'évaluation :

Les valeurs de la matrice de confusion proviennent des prévisions du modèle concernant l'ensemble des données, et à partir de ces valeurs, nous pouvons évaluer diverses métriques.

4.4.2.1 F1-Score :

La mesure F1 (ou score F1) permet d'évaluer la performance d'un modèle de classification binaire sur un ensemble de données. Elle combine deux métriques : la précision et le rappel. Le score F1 est calculé comme la moyenne harmonique de ces deux métriques. Son utilité est de fournir une évaluation synthétique de la capacité du modèle à classer correctement les exemples en deux catégories (positif/négatif). Un score F1 élevé indique que le modèle est performant à la fois pour la précision et le rappel, et il est calculé en utilisant la formule suivante [58] :

$$\text{F score} = 2 \times \frac{\text{précision} \times \text{Rappel}}{\text{précision} + \text{Rappel}}$$

4.4.2.2 Spécificité :

La spécificité d'un test représente la probabilité d'obtenir un résultat négatif chez un individu sain. En d'autres termes, elle évalue la capacité du test à identifier correctement les individus en bonne santé. Une spécificité élevée indique une réduction du nombre de faux positifs, ce qui signifie que le test est plus précis dans la détection des personnes non atteintes de la condition recherchée [59]. La spécificité du test est calculée en utilisant la formule suivante : le nombre de vrais négatifs divisé par la somme des vrais négatifs et des faux positifs. Cette formule est représentée par [60] :

$$\text{Sp} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

4.4.2.3 Sensibilité (Rappel) :

Le rappel, également connu sous le nom de sensibilité, mesure la proportion d'observations positives correctement prédites par rapport à toutes les observations positives réelles. Il est particulièrement utile lorsque le coût des faux négatifs est élevé, car il met l'accent sur la capacité du modèle à capturer toutes les instances positives, minimisant ainsi les omissions [61]. La sensibilité du test est déterminée par la proportion de vrais positifs parmi les individus malades, ce qui est calculé avec la formule suivante [60] :

$$\text{Se} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

4.4.2.4 Accuracy :

La métrique d'accuracy mesure la proportion de prédictions correctes réalisées par le modèle sur l'ensemble des prédictions effectuées. Autrement dit, elle représente le rapport entre le nombre de bonnes prédictions et le nombre total de prédictions. Pour calculer l'accuracy, on utilise les valeurs contenues dans la matrice de confusion en appliquant la formule suivante [62] :

$$\text{accuary} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

4.4.3 Rapport de classification :

Le rapport de classification est un outil d'évaluation des performances d'un modèle de classification. Il fournit des détails sur la précision, le rappel, le score F1 et le support pour chaque classe. Ce rapport peut être généré à partir de la matrice de confusion, et il présente les résultats pour chaque classe ainsi que leur moyenne, qu'elle soit pondérée ou non pondérée, selon l'option choisie.

Dans la Figure 4.2, le rapport de classification est représenté après avoir considéré deux classes : "normal" et "cancer". Pour chaque classe, la précision, le rappel et le score F1 sont calculés. De plus, une ligne supplémentaire dans ce tableau fournit la moyenne pondérée des métriques, où le poids est déterminé par le nombre total d'exemples dans chaque classe. Enfin, la colonne "support" indique le nombre d'échantillons présents pour chacune des deux classes [56].

	precision	recall	f1-score	support
0	0.50	0.50	0.50	294
1	0.53	0.53	0.53	315
accuracy			0.52	609
macro avg	0.52	0.52	0.52	609
weighted avg	0.52	0.52	0.52	609

FIGURE 4.2 – Rapport de classification.

4.5 Expériences et résultats :

4.5.1 Comparaison des performances des algorithmes sans sélection d'attributs ni prétraitement :

Le Tableau 4.2 présente les résultats de classification obtenus lorsque nous avons appliqué directement les différents algorithmes de classification (arbre de décision, k plus proches voisins, réseaux de neurones et machines à vecteurs de support) sur les bases de données génomiques du cancer, sans aucun prétraitement préalable des données ni utilisation d'algorithmes de sélection de variables.

Dans ce scénario initial, nous avons utilisé les ensembles de données brutes, avec tous les gènes (attributs) disponibles, sans retirer les redondances, gérer les valeurs manquantes ou appliquer des techniques de normalisation. De plus, nous n'avons pas cherché à réduire la dimensionnalité élevée des données en sélectionnant un sous-ensemble pertinent de gènes. L'objectif était d'évaluer les performances brutes des différents classificateurs sur ces données génomiques complexes et de haute dimension, sans aucun traitement préalable.

TABLE 4.2 – Taux de classification des classificateurs sans prétraitement ni sélection d'attributs.

Dataset	Arbre de decision	KNN	RN	SVM
Leukemia	81,81 %	86 %	95 %	90,90 %
Lung	81,96%	82%	92 %	85,24 %
MLL	81,81%	82%	95%	90,90 %
Ovarian	97,36%	95%	97%	97,36 %
SRBCT	96%	88%	100 %	100 %
Brain-tumor1	70,37%	67%	85 %	70,37 %
DLBCL	66,66%	96%	92 %	95,83 %
Lung-cancer	86,88%	82%	92 %	85,24%
Prostate-tumor	93,54%	84%	87%	87,09%

4.5.2 Comparaison des performances des algorithmes avec prétraitement mais sans sélection d'attributs :

Le Tableau 4.3 présente les résultats de classification obtenus lorsque nous avons appliqué les différents algorithmes (arbre de décision, k plus proches voisins, réseaux de neurones et machines à vecteurs de support) sur les bases de données génomiques du cancer, après avoir

effectué un prétraitement des données, mais sans utiliser les algorithmes de sélection de variables.

Dans ce deuxième scénario, nous avons d'abord appliqué des techniques de prétraitement sur les ensembles de données brutes. Cela comprend le nettoyage des données pour gérer les valeurs manquantes, et la normalisation pour mettre toutes les variables à la même échelle. Cependant, contrairement au troisième scénario (décrit plus loin), nous n'avons pas encore réduit la dimensionnalité élevée des données en sélectionnant un sous-ensemble de gènes (variables) pertinent. Tous les gènes disponibles ont été conservés à ce stade.

TABLE 4.3 – Taux de classification des classificateurs sans prétraitement des données mais avec sélection d'attributs.

Dataset	Arbre de decision	KNN	RN	SVM
Leukemia	86,36 %	86,36 %	95 %	100 %
Lung	80,32%	86,88%	89%	93%
MLL	77,27%	86,36%	91%	95 %
Ovarian	96,05%	94,73%	97%	100 %
SRBCT	92%	88%	100 %	100 %
Brain-tumor1	66,66%	81,48%	85%	96 %
DLBCL	70,83%	87,5%	92 %	92 %
Lung-cancer	85,24%	86,88%	87%	93%
Prostate-tumor	90,32%	74,19%	90%	94%

4.5.3 Comparaison des performances des algorithmes avec prétraitement mais sans sélection d'attributs :

Les tableaux 4.4 à 4.12 présentent les résultats de classification obtenus après avoir appliqué un prétraitement complet des données, suivi de l'utilisation de différentes méthodes de sélection de variables pertinentes, avant d'entraîner et d'évaluer les algorithmes de classification. Dans ce troisième scénario, le plus complet, nous avons d'abord préparé les données brutes en exécutant un prétraitement rigoureux. Cela comprend l'imputation des valeurs manquantes, ainsi que la normalisation des variables pour les ramener à la même échelle. Ces étapes sont cruciales pour garantir la qualité et la cohérence des données avant de les soumettre aux algorithmes d'apprentissage. Ensuite, nous avons appliqué trois méthodes différentes de sélection de variables : MIM (Maximum Information Maximale), MRMR (Maximum Relevance Minimum Redundancy) et JMI (Joint Mutual Information). Ces algorithmes permettent d'identifier

les gènes (attributs) les plus pertinents et informatifs pour la tâche de classification, tout en éliminant les redondances. Cela a permis de réduire considérablement la dimensionnalité initiale des données génomiques.

Pour chaque base de données et chaque méthode de sélection de variables, nous avons entraîné et évalué plusieurs algorithmes de classification populaires : les arbres de décision, les k plus proches voisins (KNN), les réseaux de neurones (RN) et les machines à vecteurs de support (SVM). Ces algorithmes couvrent différentes approches de classification et permettent d'identifier les meilleurs modèles pour ces données spécifiques.

4.5.3.1 MLL :

Le Tableau 4.4 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données MLL, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.4 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données MLL.

Alg de sélec	**Alg de class**	**10**	**30**	**50**	**80**	**100**	**130**	**150**	**200**	**300**
MRMR	Arbre	77,27	81,81	81,81	81,81	86,36	81,81	77,27	72,72	81,81
	KNN	81,81	90,90	86,36	81,81	86,36	81,81	77,27	81,81	86,36
	RN	95,45	95,45	95,45	95,45	95,45	100	100	100	100
	SVM	86,36	90,90	95,45	95,45	95,45	95,45	95,45	100	100
JMI	Arbre	77,27	81,81	81,81	81,81	86,36	81,81	77,27	72,72	81,81
	KNN	81,81	90,90	86,36	81,81	86,36	81,81	77,27	81,81	86,36
	RN	95,45	95,45	95,45	95,45	95,45	95,45	100	100	100
	SVM	86,36	90,90	95,45	95,45	100	100	100	100	100
MIM	Arbre	81,82	81,82	86,36	81,82	81,82	81,82	81,82	77,27	81,82
	KNN	81,82	90,91	86,36	81,82	86,36	81,82	77,27	81,82	86,36
	RN	83	90	90,91	92	81,82	81,82	81,82	83	87
	SVM	86,36	90,91	95,45	95,45	95,45	95,45	95,45	100	100

4.5.3.2 SRBCT :

Le Tableau 4.5 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données SRBCT, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.5 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données SRBCT.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	96	88	92	96	96	96	96	88	88
	KNN	96	96	98	100	100	100	96	100	100
	RN	100	100	100	100	100	100	100	100	100
	SVM	100	100	100	100	100	100	100	100	100
JMI	Arbre	96	88	92	96	96	96	96	88	88
	KNN	96	96	96	98,5	100	100	96	100	100
	RN	98	98	100	100	100	100	100	100	100
	SVM	100	100	100	100	100	100	100	100	100
MIM	Arbre	92	96	96	96	96	92	92	96	92
	KNN	96	96	96,5	98	100	100	96	100	100
	RN	98	98	99,5	100	100	100	100	100	76
	SVM	100	100	100	100	100	100	100	100	100

4.5.3.3 Lung :

Le Tableau 4.6 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Lung, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.6 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Lung.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	77,04	95,08	85,25	90,16	85,25	86,88	86,88	85,24	80,32
	KNN	80,32	90,16	93,44	93,44	93,44	93,44	93,44	93,44	91,80
	RN	85,24	95,08	96,72	96,72	95,08	95,08	95,08	96,72	95,08
	SVM	83,60	93,44	93,44	93,44	93,44	93,44	93,44	93,44	96,72
JMI	Arbre	77,04	90,16	85,24	90,16	85,24	86,88	86,88	85,24	80,32
	KNN	80,32	90,16	93,44	93,44	93,44	93,44	93,44	93,44	91,80
	RN	90,16	93	96,50	96,72	96,72	95,08	95,08	96,72	95,08
	SVM	83,60	93,44	93,44	93,44	93,44	93,44	93,44	93,44	96,72
MIM	Arbre	75,41	91,80	83,61	85,25	81,97	85,25	83,61	88,52	86,89
	KNN	70,49	81,97	72,13	83,61	73,77	90,16	73,77	81,97	86,89
	RN	86,89	93,44	93,44	95,08	95,08	93,44	93,44	95,08	95,08
	SVM	83,61	93,44	93,44	93,44	93,44	93,44	93,44	93,44	96,5

4.5.3.4 Ovarian :

Le Tableau 4.7 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Ovarian, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.7 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Ovarian.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	96	97,36	98,68	98,68	98,69	97,36	98,68	98,68	98,68
	KNN	98	98,68	98,68	98,68	98,68	100	100	100	100
	RN	97	97,36	100	100	100	100	100	100	100
	SVM	100	100	100	100	100	100	100	100	100
JMI	Arbre	96,05	98,68	98,68	98,68	98,68	98,68	98,68	98,68	98,68
	KNN	96	98,68	98,68	98,68	99,52	99,52	100	100	100
	RN	97	97,36	98,5	98,5	100	100	100	100	100
	SVM	100	100	100	100	100	100	100	100	100
MIM	Arbre	97,37	97,37	98,68	98,68	98,68	98,68	100	98,68	98,68
	KNN	98,68	96,05	97,37	90,79	100	96,05	98,68	97,37	94,74
	RN	98	98,68	100	100	100	100	100	100	100
	SVM	100	100	100	100	100	100	100	100	100

4.5.3.5 Leukemia :

Le Tableau 4.8 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Leukemia, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.8 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Leukemia.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	86,36	95,45	90,90	90,90	77,27	86,36	81,81	90,90	90,90
	KNN	90,90	86,36	86,36	90,90	95,45	95,45	100	100	100
	RN	86,36	86,36	90,90	90,90	90,90	90,90	100	95,45	90,90
	SVM	90,90	90,90	90,90	90,90	95,45	100	100	100	100
JMI	Arbre	86,36	86,36	90,90	90,90	90,90	90,90	86,36	86,36	81,81
	KNN	95,45	95,45	95,45	95,45	95,45	95,45	100	100	100
	RN	86,36	86,36	86,36	90,90	90,90	90,90	90,90	100	95,45
	SVM	90,90	90,90	90,90	90,90	95,45	100	100	100	100
MIM	Arbre	86,36	86,36	90,91	95,45	90,91	90,91	86,36	81,82	77,27
	KNN	86,36	90,91	90,91	90,91	95,45	95,45	100	100	100
	RN	63,64	72,73	72,73	86,36	86,36	90,91	95,45	77,27	59,09
	SVM	90,91	90,91	90,91	95,45	95,45	100	100	100	100

4.5.3.6 Brain-tumor1 :

Le Tableau 4.9 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Brain-tumor1, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.9 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Brain-tumor1.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	77,77	85,18	81,48	74,07	74,07	77,77	77,77	77,77	81,48
	KNN	77,77	77,77	81,48	70,37	70,37	70,37	70,37	70,37	74,07
	RN	81,48	92,59	92,59	92,59	88,88	88,88	85,18	85,18	92,59
	SVM	92,59	92,59	92,59	96,29	96,29	96,29	92,59	92,59	92,59
JMI	Arbre	74,07	85,18	85,18	81,48	74,07	70,37	74,07	77,77	74,07
	KNN	85,18	77,77	81,48	70,37	70,37	70,37	70,37	70,37	74,07
	RN	81,48	88,88	85,18	92,59	88,88	85,18	85,18	88,88	92,59
	SVM	92,59	92,59	92,59	96,29	96,29	96,29	92,59	92,59	92,59
MIM	Arbre	77,78	81,48	85,19	74,07	74,07	81,48	70,37	77,78	77,78
	KNN	85,19	77,78	81,48	70,37	70,37	70,37	70,37	70,37	74,07
	RN	51,85	81,48	92,59	88,89	70,37	77,78	74,07	74,07	81,48
	SVM	92,59	92,59	92,59	96,30	96,30	96,30	92,59	92,59	92,59

4.5.3.7 DLBCL :

Le Tableau 4.10 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données DLBCL, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.10 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données DLBCL.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	70,83	75	70,83	75	70,83	66,66	70,83	66,66	58,33
	KNN	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83
	RN	91,66	91,66	95,83	95,83	100	100	95,83	100	100
	SVM	95,83	95,83	95,83	100	100	100	95,83	95,83	100
JMI	Arbre	70,83	70,83	70,83	75	75	66,66	66,66	70,83	75
	KNN	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83
	RN	95,83	91,66	95,83	100	100	91,66	87,50	95,83	87,50
	SVM	95,83	95,83	100	100	100	95,83	95,83	95,83	100
MIM	Arbre	66,67	70,83	70,83	70,83	70,83	70,83	70,83	70,83	70,83
	KNN	62,50	87,50	45,83	50	75	66,67	54,17	75	83,33
	RN	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83	95,83
	SVM	95,83	95,83	95,83	100	100	95,83	95,83	95,83	100

4.5.3.8 Lung-cancer :

Le Tableau 4.11 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Lung-cancer, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.11 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Lung-cancer.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	77,04	93,44	83,60	85,24	88,52	85,24	85,24	86,88	75,40
	KNN	80,32	90,16	93,44	93,44	93,44	93,44	93,44	93,44	91,80
	RN	85,24	96,72	96,72	95,08	95,08	95,08	95,08	95,08	95,08
	SVM	83,60	93,44	93,44	93,44	93,44	93,44	93,44	93,44	95,82
JMI	Arbre	77,04	95,08	88,52	83,60	81,96	80,32	85,24	83,60	86,88
	KNN	80,32	90,16	93,44	93,44	93,44	93,44	93,44	93,44	91,80
	RN	88,52	88,52	95,08	96,72	95,08	93,44	95,08	95,08	95,08
	SVM	83,60	93,44	93,44	93,44	93,44	93,44	93,44	93,44	95,44
MIM	Arbre	75,41	86,89	85,25	90,16	91,80	86,89	81,97	91,80	81,97
	KNN	80,33	90,16	93,44	93,44	93,44	93,44	93,44	93,44	91,80
	RN	81,97	59,02	77,05	77,05	90,16	80,33	80,33	80,33	98,36
	SVM	83,61	93,44	93,44	93,44	93,44	93,44	93,44	93,44	93,44

4.5.3.9 Prostate-tumor :

Le Tableau 4.12 représente les taux de classification obtenus après avoir appliqué les étapes mentionnées précédemment sur la base de données Prostate-tumor, avec différents nombres de caractéristiques sélectionnées.

TABLE 4.12 – Taux de classification obtenus après prétraitement et sélection d'attributs sur le jeu de données Prostate-tumor.

Alg de sélec	Alg de class	10	30	50	80	100	130	150	200	300
MRMR	Arbre	90,32	80,64	83,87	87,09	90,32	87,09	96,77	80,64	83,87
	KNN	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54
	RN	90,32	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54
	SVM	93,54	93,54	93,54	93,54	93,54	93,54	93,54	95,54	95,54
JMI	Arbre	87,09	77,41	80,64	80,64	87,09	93,54	93,54	90,32	90,32
	KNN	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54
	RN	90,32	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54
	SVM	93,54	93,54	93,54	93,54	93,54	93,54	93,54	95,54	95,54
MIM	Arbre	90,32	87,10	83,87	87,10	87,10	90,32	83,87	90,32	80,65
	KNN	93,55	93,55	93,55	93,55	93,55	93,55	93,55	93,55	93,55
	RN	70,97	70,97	83,87	93,55	83,87	93,55	83,87	77,42	93,55
	SVM	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54	93,54

4.6 Comparaison des mesures de performance (accuracy, rappel, F1-score et précision) avant et après la sélection :

Nous avons calculé les mesures de performance suivantes : la précision, le rappel, le score F1 et l'exactitude pour différents algorithmes de classification, notamment les arbres de décision, les k-plus proches voisins (KNN), les réseaux de neurones (NN) et les machines à vecteurs de support (SVM). Ces mesures ont été évaluées dans deux scénarios distincts : le premier où nous avons appliqué le prétraitement des données sans effectuer de sélection de variables, et le second où nous avons combiné le prétraitement avec l'application d'algorithmes de sélection des variables pertinentes.

L'objectif était de comparer les performances des différents algorithmes dans ces deux contextes, avec ou sans réduction de dimensionnalité par sélection de variables, afin d'identifier les approches les plus appropriées pour la classification des données génomiques du cancer. Auparavant, nous avions comparé les algorithmes uniquement sur l'exactitude, ce qui peut donner une vision partielle des performances. Certains modèles peuvent avoir une exactitude élevée mais détecter mal une des classes (par exemple, un taux élevé de faux négatifs). En ajoutant les mesures de précision, de rappel et de score F1, nous obtenons un aperçu plus complet des forces et des faiblesses de chaque algorithme, permettant une meilleure comparaison.

4.6.1 Comparaison des mesures de performance (accuracy, rappel, F1-score et précision) avant la sélection :

Le tableau présenté dans l'image 4.3 illustre les résultats en termes d'exactitude, de score F1, de précision et de rappel obtenus après avoir appliqué les différents algorithmes de classification mentionnés précédemment, à savoir les arbres de décision, les k-plus proches voisins (KNN), les réseaux de neurones (RN) et les machines à vecteurs de support (SVM), sur les bases de données. Cependant, ces résultats ont été obtenus après un prétraitement des données, mais sans appliquer d'algorithmes de sélection de variables.

Algorithm de classification	Arbre de decision				KNN				RN				SVM			
Mesures de performance / Dataset	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel
MLL	72,72	72,69	73,21	72,61	86,36	86,79	87,83	86,30	91	90	90	90	95	95	96	95
Lung	88,52	68,90	71,75	66,50	86,88	81,96	95,83	75,83	89	84	85	85	93	92	98	88
Leukemia	90,90	84,72	84,72	84,72	86,36	66,15	92,85	62,5	95	93	90	97	100	100	100	100
Lung-cancer	88,52	68,90	71,75	66,50	86,88	81,96	95,83	75,83	87	73	76	71	93	92	98	88
SRBCT	96	96,06	95,83	96,87	88	88,74	90,35	90,62	100	100	100	100	100	100	100	100
Brain-tumor1	66,66	42,68	45,33	40,55	81,48	46,89	55,65	43,33	85	65	58	77	96	79	99	80
Ovarian	98,68	98,49	98	99,03	94,73	93,91	93,91	93,91	97	97	97	97	100	100	100	100
DLBCL	66,66	40	40	40	87,50	79,48	77,36	82,50	92	85	85	85	92	85	85	85
Prostate-tumor	90,32	89,94	90,57	89,52	74,19	74,16	77,41	76,70	90	90	90	91	94	93	93	93

FIGURE 4.3 – Mesures de performances des classificateurs sans sélection d'attributs.

4.6.2 Comparaison des mesures de performance (accuracy, rappel, F1-score et précision) après la sélection :

Les tableaux exposés dans les images 4.4, 4.5 et 4.6 mettent en lumière les résultats concernant l'exactitude, le score F1, la précision et le rappel obtenus après l'application des divers algorithmes de classification mentionnés précédemment : les arbres de décision, les k-plus proches voisins (KNN), les réseaux de neurones (RN) et les machines à vecteurs de support (SVM). Ces résultats ont été obtenus suite à un prétraitement des données, suivi de l'application d'algorithmes de sélection de variables (MRMR, MIM, JMI) avec différents nombres d'attributs sélectionnés. Ces tableaux permettent d'évaluer l'impact de la sélection de variables pertinentes, combinée au prétraitement, sur les performances de classification en fonction du nombre de caractéristiques retenues.

Il est important de noter que le tableau 4.4 contient les résultats avec l'algorithme de sélection MRMR, tandis que le tableau 4.5 contient les résultats avec l'algorithme de sélection MIM, et enfin le tableau 4.6 contient les résultats avec l'algorithme de sélection JMI.

Algorithm de classification	Arbre de decision				KNN				RN				SVM				
Mesures de performance / Dataset	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Nombre d'attributs Selectionne -es
MLL	82	82	82	82	95	96	96	96	100	100	100	100	100	100	100	100	300
Lung	82	61	63	60	92	91	96	82	95	94	98	91	94	93	98	90	100
Leukemia	85	84	82	90	89	83	92	77	97	95	92	98	100	100	100	100	150
Lung-cancer	82	64	71	59	90	87	97	82	97	96	99	92	93	92	97	88	80
SRBCT	94	90	96	89	88	71	69	75	100	100	100	100	100	100	100	100	200
Brain-tumor1	75	60	63	60	89	56	57	55	89	86	81	97	97	79	99	80	100
Ovarian	97	96	96	97	98,5	98	97	97	99	99	98,5	99,25	100	100	100	100	250
DLBCL	67	49	50	50	92	85	85	85	96	93	90	97	100	100	100	100	80
Prostate-tumor	78	78	84	82	86	87	87	88	90	90	90	91	98	95,5	95	95	50

FIGURE 4.4 – Mesures de performances de MRMR avec un classificateur Arbre de décision, KNN, RN, SVM.

Algorithm de classification	Arbre de decision				KNN				RN				SVM				
Mesures de performance / Dataset	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 1- score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Nombre d'attributs Selectionne -es
MLL	86	86	86	86	86	86	88	87	95	95	96	95	100	100	100	100	200
Lung	79	59	61	58	82	69	70	70	93	91	98	89	93	91	98	89	300
Leukemia	82	74	72	79	91	81	95	75	95	93	90	97	100	100	100	100	150
Lung-cancer	85	65	67	65	90	87	96	82	79	67	65	71	93	92	97	88	80
SRBCT	88	93	93	92	84	69	67	72	90	91	94	91	100	100	100	100	250
Brain-tumor1	74	54	63	62	89	56	57	55	93	77	76	79	96	81	81	80	80
Ovarian	95	95	95	95	96	96	97	96	97	97	98	97	100	100	100	100	250
DLBCL	71	41	40	42	92	85	85	85	71	66	68	82	100	100	100	100	80
Prostate-tumor	84	84	86	86	87	87	87	88	84	83	84	83	95	94	94	94	50

FIGURE 4.5 – Mesures de performances de MIM avec un classificateur Arbre de décision, KNN, RN, SVM.

Algorithm de classification	Arbre de decision				KNN				RN				SVM				
Mesures de performance / Dataset	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Accuracy	F1 - score	precision	Rappel	Nombre d'attributs Selectionne -es
MLL	86	86	86	86	86	86	88	87	95	95	96	95	100	100	100	100	200
Lung	79	59	61	58	82	69	70	70	93	92	98	89	92	91	95	88	300
Leukemia	82	74	72	79	91	81	95	75	95	93	90	97	100	100	100	100	150
Lung-cancer	80	59	64	55	90	87	97	82	97	95	99	93	93	92	98	88	80
SRBCT	88	93	93	92	84	69	67	72	90	91	94	91	100	100	100	100	250
Brain-tumor1	74	54	63	62	89	56	57	55	93	77	76	80	96	81	81	80	80
Ovarian	95	95	95	95	96	96	97	96	97	97	98	97	100	100	100	100	250
DLBCL	67	49	50	50	92	75	75	75	96	93	90	97	100	100	100	100	80
Prostate-tumor	84	84	86	86	87	87	87	88	84	83	84	83	95	94	94	94	50

FIGURE 4.6 – Mesures de performances de JMI avec un classificateur Arbre de décision, KNN, RN, SVM.

4.7 Discussion des résultats

Dans cette section, nous présentons et analysons les résultats obtenus en utilisant différents algorithmes de sélection : Mutual Information Maximization (MIM), Maximum Relevance Minimum Redundancy (MRMR), Joint Mutual Information (JMI), et des algorithmes de classification : arbre de décision, k-plus proches voisins, réseaux de neurones, et machines à vecteurs de support.

Nous avons d'abord comparé les performances des algorithmes de classification en nous basant sur le taux de classification selon trois scénarios. Les résultats sont présentés dans les tableaux 4.2 à 4.16 :

- **Scénario 1 :** Application directe des algorithmes de classification sur les bases de données, sans prétraitement ni sélection des variables. Ce scénario nous a permis de poser une base de référence pour mesurer l'impact du prétraitement et des étapes de sélection des variables sur l'amélioration des performances de classification.
- **Scénario 2 :**Application des algorithmes de classification après prétraitement des don-

nées mais sans sélection des variables. Ce tableau nous a permis d'évaluer l'impact du prétraitement des données seul sur les performances de classification, sans sélection de variables.

- **Scénario 3 :**Application des algorithmes de classification après prétraitement et sélection des variables. Ce scénario nous a permis de mesurer l'impact combiné du prétraitement et de la sélection des variables sur l'amélioration des performances de classification.

Ensuite, nous avons évalué les performances avec des métriques plus complètes : l'accuracy, le F1-score, la précision et le rappel. À cette étape, seuls les scénarios 2 et 3 ont été considérés, car se baser uniquement sur le taux de classification peut donner une vision partielle. Les résultats sont illustrés dans les figures 4.3 à 4.6.

1. Concernant la comparaison basée sur le taux de classification, les résultats obtenus sont illustrés dans les tableaux suivants :

   - Le tableau 4.2 présente les résultats du premier scénario sans prétraitement ni sélection de variables : les performances étaient très faibles. En les comparant, on constate que l'algorithme des réseaux de neurones (RN) était le plus efficace avec la majorité des bases de données utilisées, suivi de l'algorithme des machines à vecteurs de support (SVM) qui offrait de meilleures performances avec certaines bases comme ovarian et DLBCL.

   - Le tableau 4.3 présente les résultats du deuxième scénario avec prétraitement mais sans sélection de variables : les résultats étaient meilleurs que ceux du tableau 4.2. Après prétraitement, l'algorithme SVM présentait les meilleures performances avec toutes les bases de données.

   - Les tableaux de 4.4 à 4.14 correspondent au scénario 3 : nous avons effectué la sélection avec différents nombres de variables (10, 30, 50, 80, 100, 130, 150, 200, 300). Les performances sont encore améliorées, atteignant jusqu'à 100 % avec les machines à vecteurs de support et différents nombres de variables sélectionnées selon les algorithmes.
2. Concernant la comparaison basée sur l'accuracy, le F1-score, la précision et le rappel, les résultats obtenus sont illustrés dans les tableaux suivants :

- Dans le scénario 2 (figure 4.3), SVM avait les meilleures performances sur toutes les bases. RN était comparable à SVM sur SRBCT et DLBCL.

- Dans le scénario 3 avec mRMR (figure 4.4), les résultats étaient bien meilleurs qu'au scénario 2, validant l'apport de la sélection de variables. SVM était le meilleur, sauf sur Lung et Lung-cancer où RN était préférable. Sur MLL et SRBCT, RN et SVM avaient des performances équivalentes.

- Avec MIM (figure 4.5) et JMI (figure 4.6), les conclusions étaient similaires : nette amélioration grâce à la sélection, et SVM présentait généralement les meilleures performances, excepté sur Lung avec JMI où RN était meilleur.

En résumé, l'utilisation des SVM avec l'algorithme de sélection MIM s'est avérée être la meilleure approche pour obtenir une classification précise et robuste des ensembles de données étudiés, car elle a permis d'obtenir un taux de classification de 100 % sur la majorité des bases de données après sélection des variables pertinentes. Le classificateur de réseaux de neurones (RN) représente une alternative intéressante. Les algorithmes d'arbre de décision et de k-plus proches voisins (KNN) étaient généralement moins efficaces. L'application d'étapes de prétraitement des données, notamment la sélection de variables à l'aide de MIM, mRMR et JMI, a permis d'améliorer considérablement la précision de tous les modèles. En fin de compte, SVM combiné à la technique de sélection de variables constitue le meilleur choix pour obtenir les résultats de classification les plus performants sur les ensembles de données étudiés, avec les RN comme alternative solide.

4.8 Conclusion

Ce chapitre a présenté une étude comparative rigoureuse des performances de différents algorithmes de sélection d'attributs et de classification sur des données génomiques de cancers réelles. Une méthodologie solide, utilisant Python et ses bibliothèques, a permis de mener des expériences selon divers scénarios de prétraitement et sélection d'attributs. L'évaluation détaillée des modèles a révélé que certains algorithmes comme SVM et les réseaux de neurones offrent d'excellentes performances après sélection pertinente des variables, tandis que d'autres comme les arbres de décision sont moins performants. L'importance cruciale des étapes de prétraitement et de sélection d'attributs a été soulignée. Ces conclusions ouvrent la voie à l'optimi-

sation des meilleurs algorithmes, au développement d'approches hybrides et à leur application potentielle en oncologie pour le diagnostic précoce personnalisé des cancers.

Conclusion générale

Dans cette étude, notre objectif principal était de relever le défi crucial consistant à identifier les gènes les plus informatifs et pertinents, pour une détection précise et fiable des divers types de cancers. Pour y parvenir, nous avons élaboré un processus structuré en trois étapes, chacune visant à évaluer l'efficacité des algorithmes de classification dans différentes conditions.

Dans un premier temps, nous avons appliqué directement les algorithmes de classification sur les données brutes, sans aucune manipulation préalable. Ensuite, nous avons amélioré la qualité des données en les prétraitant, notamment par la normalisation, la gestion des valeurs manquantes et la réduction du bruit, avant de réappliquer les mêmes algorithmes. Enfin, nous avons optimisé les données prétraitées en sélectionnant les gènes les plus pertinents à l'aide de techniques spécifiques avant de les soumettre aux algorithmes de classification.

Les résultats obtenus ont montré que l'algorithme des machines à vecteurs de support (SVM) s'est distingué comme le plus performant, atteignant un taux de classification de 100% sur la majorité des bases de données après la sélection des variables pertinentes. Les réseaux de neurones (RN) ont également présenté des performances prometteuses. En revanche, les performances des algorithmes des arbres de décision et des k-plus proches voisins (KNN) étaient globalement inférieures.

La solution proposée, qui implique une comparaison systématique d'algorithmes à travers différents scénarios de prétraitement et de sélection de variables, présente plusieurs avantages notables. Tout d'abord, elle permet une évaluation exhaustive des performances des divers algorithmes de classification sur des données génomiques, offrant ainsi un aperçu complet de leurs forces et faiblesses respectives. De plus, en intégrant des étapes de prétraitement et de sélection de variables, cette approche contribue à améliorer la qualité des données d'entrée en réduisant le bruit et la redondance, ce qui augmente la précision des modèles résultants.

Cependant, malgré ses nombreux avantages, la solution proposée présente également certaines limites à considérer. L'une des principales contraintes réside dans la complexité computationnelle accrue résultant de l'application de multiples algorithmes et techniques de prétraitement sur de vastes ensembles de données génomiques. De plus, bien que les techniques de sélection de variables telles que MIM, mRMR et JMI se soient révélées efficaces, elles peuvent parfois ne pas saisir certaines interactions complexes entre les gènes, nécessitant ainsi l'exploration de méthodes alternatives ou hybrides. Enfin, étant donné que notre étude s'est concentrée sur des ensembles de données spécifiques, la généralisation de nos résultats à d'autres contextes pourrait être limitée.

Dans le cadre de travaux scientifiques futurs, nous prévoyons d'explorer des approches hybrides combinant différentes techniques de sélection de gènes. Cette démarche vise à tirer pleinement parti de la richesse des données génomiques disponibles tout en améliorant encore la précision de la détection du cancer. Cette avenue de recherche prometteuse ouvre de nouvelles perspectives pour une détection plus précoce et plus précise des divers types de cancers. En développant des techniques combinant différentes approches de sélection de gènes, nous espérons contribuer à l'amélioration des résultats cliniques et à une meilleure qualité de vie pour les patients atteints de cancer.

Références

[1] B. HAMZA, "contribution à la modélisation et la simulation monte carlo pour l'optimisation du traitement en curiethérapie à haut débit de dose : applications aux traitements des cancers localisés," 2020.

[2] P. J. e. F. P. A. Stratton, Michael R et Campbell, "Le génome du cancer," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[3] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer : the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[4] A. Rammal, "Modélisation multi-agent dans un processus de gestion multi-acteur, application au maintien à domicile," Thèse de doctorat en Informatique, Université Toulouse 3, Toulouse, France, 2010.

[5] Y. Slimani, M. A. Essegir, M. L. Samb, F. Camara, and S. Ndiaye, "Approche de sélection d'attributs pour la classification basée sur l'algorithme rfe-svm," *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées*, vol. 17, 2014.

[6] N. Bidi, "Approches méta-heuristiques pour les tâches de classification," Thèse de doctorat en Informatique, Université Djillali Liabes Sidi Bel Abbès, Sidi Bel Abbès, Algérie, 2018.

[7] R. R. Lallich, "Construction d'arbres de décision par optimisation."

[8] S. Amine, C. Dartigues-Pallez, and R. Gaetan, "Classification supervisée de données pédagogiques pour la réussite dans l'enseignement supérieur," Ph.D. dissertation, I3S, Université Côte d'Azur, 2020.

[9] L. Ouchiha, "Classification supervisée de documents : étude comparative," Essai de

deuxième cycle, Université du Québec en Outaouais, Gatineau, Canada, 2016.

[10] C. Tissot, “Les techniques bayésiennes appliquées à des données d’accidentologie,” in *Congrès Lambda Mu 22 «Les risques au cœur des transitions»(e-congrès)-22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*, 2020.

[11] O. Moussati, “Classification des données de biopuces : étude comparative,” Mémoire de master en Informatique, Université des Sciences et de la Technologie d’Oran Mohamed Boudiaf, Oran, Algérie, 2016.

[12] N. Le Meur, “De l’acquisition des données de puces à adn vers leur interprétation : Importance du traitement des données primaires,” Thèse de doctorat en Bio-informatique, Université de Nantes, Nantes, France, 2005.

[13] S. Ferré, “Concepts de plus proches voisins dans des graphes de connaissances.” in *28es Journées francophones d’Ingénierie des Connaissances IC 2017*, 2017, pp. 163–174.

[14] M. Ouguissi and A. Almaoui, “Prédiction de rayonnement solaire journalière par réseau de neurone,” Mémoire de master en Informatique, Université Kasdi Merbah Ouargla, Ouargla, Algérie, 2020.

[15] N. Gueye, “Exploration des liens formels entre les méthodes statistiques et neuronales en classification,” Mémoire de maîtrise en Mathématiques et informatique appliquées, Université du Québec à Trois-Rivières, Trois-Rivières, Canada, 2019.

[16] S. ROUANE, N. elhouda ROUANE, and M. S. B. YAHIA, “Indentification par reseaux de neurone.”

[17] J. C. H. Hernandez, “Algorithmes métaheuristiques hybrides pour la sélection de gènes et la classification de données de biopuces,” Ph.D. dissertation, Université d’Angers, 2008.

[18] G. Lebrun, “Sélection de modèles pour la classification supervisée avec des svm (séparateurs à vaste marge). application en traitement et analyse d’images.” Ph.D. dissertation, Université de Caen Basse-Normandie, 2006.

[19] S. El Ferchichi, “Sélection et extraction d’attributs pour les problèmes de classification,”

Thèse de doctorat, UNIVERSITÉ Lille 1, Lille, France, 2013. [Online]. Available : https://www.theses.fr/2013LIL10042

[20] M. Z. Nawel, "Techniques d'apprentissage pour la sélection d'attributs : Application à la reconnaissance des formes," Ph.D. dissertation, BADJI MOKHTAR UNIVERSITY, 2018.

[21] E. B. Huerta, "Logique floue et algorithmes génétiques pour le pré-traitement de données de biopuces et la sélection de gènes," Ph.D. dissertation, Université d'Angers, 2008.

[22] A. BOUBLENZA, "Coopération entre classifieurs hétérogènes pour la reconnaissance des données médicales," Ph.D. dissertation, UNIVERSITE ABOU-BEKR BELKAID – TLEMCEN, 2017.

[23] F. Mhamdi and M. Kchouk, "Algorithme hybride de sélection d'attributs pour le classement des protéines," *EGC 2014*.

[24] A. T. E. SELMI, "Pré-traitement de données de biopuces et la sélection de gènes par une approche bioinspirée."

[25] M. M. Kamilia, "Approches bio-inspirées pour la sélection d'attributs," Ph.D. dissertation, Université Badji Mokhtar-Annaba, 1945.

[26] N. Challita, "Contributions à la sélection des attributs de signaux non stationnaires pour la classification," Ph.D. dissertation, Université de Technologie de Troyes, 2018.

[27] R. Bendana, "Sélection d'attributs basée sur un algorithme génétique neuronal : Application à la reconnaissance des caractères manuscrits," Master's thesis, Université Mentouri de Constantine, 2007.

[28] A. Alaoui *et al.*, "Hybridation des métaheuristiques dans le processus d'extraction de connaissances à partir de données." Ph.D. dissertation, 2021.

[29] A. Abdiya, "Application des techniques des métaheuristiques pour l'optimisation de la tâche de la classification de la fouille de données," Master's thesis, UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN Mohamed Boudiaf, 2012.

[30] M. Kalakech, "Sélection semi-supervisée d'attributs : application à la classification de

textures couleur," Ph.D. dissertation, Lille 1, 2011.

[31] D. Dernoncourt, "Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique," Ph.D. dissertation, Université Pierre et Marie Curie-Paris VI, 2014.

[32] H. CHOUAIB, "Sélection de caractéristiques : méthodes et applications," Ph.D. dissertation, Université Paris Descartes, 2011.

[33] H. Mébaki, "Une méthode hybride basée sur l'information mutuelle et les algorithmes génétiques pour la sélection des attributs," 2023.

[34] V. T. Kishorbhai and M. Malaviya, "Overview of an ovarian cancer and its treatment aspects," vol. 9, 2021.

[35] B. Kumar, "Modalities of diagnosis & management of small blue round cell tumor," *International Journal of Medicine and Pharmaceutical Science (IJMPS) ISSN (P)*, pp. 2250–0049.

[36] A. Davis, A. Viera, and M. Mead, "Leukemia : an overview for primary care." *American family physician*, vol. 89 9, pp. 731–8, 2014.

[37] R. Kurman and I. Shih, "The origin and pathogenesis of epithelial ovarian cancer : A proposed unifying theory," *The American Journal of Surgical Pathology*, vol. 34, pp. 433–443, 2010.

[38] R. Stam, M. L. Boer, M. Passier, S. Sallan, S. Armstrong, and P. Pieters, "Mll rearranged infant acute lymphoblastic leukemia is characterized by silencing of the putative tumor suppressor gene fhit." *Blood*, vol. 104, pp. 525–525, 2004.

[39] R. A. Taylor, R. Toivanen, and G. Risbridger, "Stem cells in prostate cancer : treating the root of the problem." *Endocrine-related cancer*, vol. 17 4, pp. R273–85, 2010.

[40] P. Hoffman, A. Mauer, and E. Vokes, "Lung cancer," *The Lancet*, vol. 355, pp. 479–485, 2000.

[41] E. S. Jaffe, *Pathology and genetics of tumours of haematopoietic and lymphoid tissues*. Iarc, 2001, vol. 3.

[42] A. Batta, “Increasing incidence of brain tumors,” vol. 4, pp. 13–21, 2017.

[43] A. Kusiak, “Entropie de l’information,” 2006.

[44] G. Eric and Y. François, *Modèles statistiques pour l’accès à l’information textuelle*. Lavoisier, 2011.

[45] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[46] M. Rennela, “Causalité informationnelle : l’information est-elle un bon fondement pour la physique quantique ?”

[47] A. Adjimi, “Réduction de la dimensionnalité pour l’amélioration des performances d’identification biométrique.” Ph.D. dissertation, Université de Bordj Bou Arréridj-Mohamed Bachir El Ibrahimi, 2018.

[48] H. Daviet, “Class-add, une procédure de sélection de variables basée sur une troncature k-additive de l’information mutuelle et sur une classification ascendante hiérarchique en pré-traitement,” Ph.D. dissertation, Université de Nantes, 2009.

[49] X. Gu, J. Guo, L. Xiao, T. Ming, and C. Li, “A feature selection algorithm based on equal interval division and minimal-redundancy–maximal-relevance,” *Neural Processing Letters*, vol. 51, pp. 1237–1263, 2019.

[50] Z. Zeng and X. Heng, “Feature selection and visualization based on interaction dominance,” *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pp. 668–673, 2019.

[51] T. Li, X. Chen, S. Zhang, Z. Dong, and K. Keutzer, “Cross-domain sentiment classification with contrastive learning and mutual information maximization,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8203–8207.

[52] F. Nelli, *Python data analytics : Data analysis and science using PANDAs, Matplotlib and the Python Programming Language*. Apress, 2015.

[53] L. Dekkiche, “Classification des arythmies ecg avec des méthodes de machine learning et de deep learning.” Ph.D. dissertation, Université Mouloud Mammeri, 2020.

[54] F. Nelli, “The numpy library,” pp. 35–61, 2015.

[55] N. Kwak and C.-H. Choi, “Input feature selection by mutual information based on parzen window,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 1667–1671, 2002.

[56] N. E. H. Derouaz and A. Kedjouti, “Proposition d’une technique efficace pour l’analyse du big data basée sur l’intelligence artificielle,” Master’s thesis, Université de Mohamed El Bachir El Ibrahimi, 2023.

[57] I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, “Multi-class confusion matrix reduction method and its application on net promoter score classification problem,” in *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, 2021, pp. 412–419.

[58] A. LATI *et al.*, “Etude statistique pour le choix des paramètres de classification des images médicale par des modèles d’intelligence artificielle,” Ph.D. dissertation, Université Kasdi Merbah Ouargla.

[59] D. Bertrand, J. Fluss, C. Billard, and J. C. Ziegler, “Efficacité, sensibilité, spécificité : Comparaison de différents tests de lecture,” *L’Année psychologique*, vol. 110, no. 2, pp. 299–320, 2010.

[60] H. Delacour, A. Servonnet, A. Perrot, J. Vigezzi, and J. Ramirez, “La courbe roc (receiver operating characteristic) : principes et principales applications en biologie clinique,” in *Annales de biologie clinique*, vol. 63, no. 2, 2005, pp. 145–154.

[61] L. Matthews, *ÉMERGENCE : L’AUBE DE L’INTELLIGENCE ARTIFICIELLE CONSCIENTE*. Larry Matthews, 2023. [Online]. Available : https://books.google.dz/books?id=QUvhEAAAQBAJ

[62] A. LATI *et al.*, “Etude statistique pour le choix des paramètres de classification des images médicale par des modèles d’intelligence artificielle,” Ph.D. dissertation, Université Kasdi Merbah Ouargla.