وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR UNIVERSITY – ANNABA

UNIVERSITE BADJI MOKHTAR – ANNABA

جامعة باجي مختار – عنابة

Faculté: Sciences de l'Ingéniorat

Département: Informatique

THÈSE

Présentée en vue de l'obtention du diplôme de

DOCTORAT 3ème cycle

Intitulée

Paradigmes Avancés de l'Apprentissage Automatique pour l'Analyse et la Classification des Données Multimodales

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : STIC (**S**ciences et **T**echnologies de l'**I**nformation et de la **C**ommunication)

Par

Mr. Nacer Eddine BENZEBOUCHI

Devant le jury

Pr. Nabiha AZIZI	**Université Badji Mokhtar – Annaba**	**Directeur**
Pr. Hassina SERIDI	**Université Badji Mokhtar – Annaba**	**Président**
Pr. Nadir FARAH	**Université Badji Mokhtar – Annaba**	**Examinateur**
Dr. Brahim FAROU	**Université 8 mai 1945 – Guelma**	**Examinateur**

Année 2019/2020

Remerciements

Je tiens à remercier, le bon dieu, le tout puissant et miséricordieux, qui m'a donné la force, la foi et la patience d'accomplir ce travail.

En second lieu, je tiens à remercier vivement mon directeur de thèse Madame le Professeur *Nabiha AZIZI* pour son précieux conseil, ses riches orientations, son soutien et son aide durant toute la période pendant laquelle elle m'a accordé toute l'attention nécessaire pour l'accomplissement de mon travail. Également, je voudrais lui signifier ma gratitude et mes sincères salutations.

J'exprime également toute ma reconnaissance au Professeur *Hassina SERIDI* pour avoir aimablement présidé le jury de cette thèse. En outre, je tiens à présenter mes sincères remerciements au Professeur *Nadir FARAH* et au Docteur *Brahim FAROU* de l'honneur qu'ils m'ont accordé de participer au jury de ma thèse. D'une manière générale, je remercie tous les membres du jury pour l'intérêt qu'ils ont fourni en vue de finaliser ma recherche par l'acceptation et l'enrichissement de l'étude et de l'examen mon travail grâce à leurs propositions adéquates. Ma profonde reconnaissance envers le Docteur *Didier SCHWAB* du Laboratoire d'Informatique de Grenoble pour sa collaboration et serviabilité qu'il m'a témoignée au cours de ces dernières années.

Bien sûr, mes vifs remerciements sont adressés à ma très chère famille: *mon père*, *ma mère*, *mes frères* et *ma seule sœur*. Je vous remercie chaleureusement de votre soutien ainsi que les encouragements que vous avez prodigués tout au long de la réalisation de cette thèse. De manière concrète, je les remercie pour l'intérêt constant porté à mes travaux. Je suis aussi conscient de l'héritage culturel et intellectuel reçu au cours de mon éducation et qui a eu une influence sur les nombreux choix effectués pendant la thèse. Pour cela, je remercie *mes parents*.

Par ailleurs, je présente mes remerciements sincères à tous les membres de laboratoire *LabGED*.

Enfin, je tiens à remercier également l'ensemble des personnes qui ont participé de près ou de loin à la réalisation de ce travail.

ملخص

أصبح التصنيف متعدد الوسائط مجالًا نشطًا جدًا للبحث في السنوات الأخيرة، في مجال التعلم الآلي والذكاء الاصطناعي بسبب نتائجه المشرفة في مختلف التطبيقات ذات الأهمية العملية التي تسمح بالتكامل والجمع كل هذه الطرائق/الأشكال من طبائع مختلفة في كل متماسك، في تجربة العولمة.

تركز هذه الأطروحة على تحليل تأثير واستخدام التعلم متعدد الوسائط المطبق على العديد من مصادر المعلومات من جهة (البيانات النصية وتلك المستمدة من التصوير الطبي) و من ناحية أخرى، وفقًا لطبيعة القواعد المعالجة (متوازنة أو غير متوازنة) من خلال اعتماد تقنيات متقدمة للتعلم الآلي، ولا سيما التعلم العميق بالإضافة إلى المناهج الشاملة/المجموعة من أجل إنشاء أنظمة قوية لدعم القرار.

في المقام الأول، نهج متعدد الوسائط يعتمد على الشبكات العصبية التلافيفية (*CNN*) ومصنف *SVM* لتصنيف / التحقق من نصوص المؤلفين وفقًا لأسلوب كتابتهم باستخدام "تضمين الكلمات" المعجمي؛ نقترح كذلك، في سياق التصوير الطبي، طريقة جديدة بعنوان: دمج المصنفات متعددة الوسائط القائمة على البيانات والتعاون المميزات/الخصائص المختلفة لتشخيص الجلوكوما من خلال تطبيق نهج اندماج متعدد الوسائط جديد ، يسمى النهج الهجين من أجل الاستفادة من اثنين من تقنيات الاندماج المبكر والمتأخر.

في المقام الثاني نتعامل مع تقنيات التجميع المتقدمة، من خلال اقتراح نهج متعدد الوسائط للتعلم يعتمد على المصنف التراص(*Stacking*) للتشخيص التلقائي لمرض السكري في سياق البيانات غير المتوازنة.

تؤكد النتائج التجريبية أن دمج الوسائط المتعددة في المساهمات المختلفة باستخدام نماذج متقدمة للتعلم الآلي قادر على توليد استنتاجات أكثر صلابة، أو حتى أفكارًا جديدة، والتي ستكون مستحيلة في نظام أحادي الوسائط، مما يسمح لنا بالاستمرار في هذا الاتجاه من البحث.

الكلمات المفتاحية: التصنيف متعدد الوسائط، الشبكات العصبية التلافيفية (CNN)، مجموعة المصنفات، التصنيف التلوي، البيانات غير المتوازنة.

Résumé

La classification multimodale est devenue un champ de recherche très actif durant ces dernières années, dans le domaine de l'apprentissage automatique et de l'intelligence artificielle en raison de ses résultats honorables dans diverses applications d'intérêt pratique permettant d'intégrer et de combiner toutes ces modalités de natures diverses en un tout cohérent, en une expérience globalisante.

Cette thèse se focalise sur l'analyse de l'impact et de l'utilisation de l'apprentissage multimodal appliquée à de nombreuses sources d'informations d'une part (les données textuelles et celles issues de l'imagerie médicale) et d'autre part selon la nature des bases traitées (équilibrées ou non) en adoptant des techniques avancées de l'apprentissage automatique, en particulier l'apprentissage profond ainsi que des approches ensemblistes afin de générer des systèmes robustes pour l'aide à la décision.

Dans un premier lieu, une approche multimodale s'appuyant sur les réseaux de neurones convolutionnels (CNN) et le classifieur SVM pour la classification/vérification des textes des auteurs selon leur style d'écriture en utilisant le plongement lexical « Word Embedding »; nous suggérons en outre, dans le contexte de l'imagerie médicale, une nouvelle méthode intitulée: la fusion des classifieurs basée données multimodales et coopération des caractéristiques pour le diagnostic du glaucome en appliquant une nouvelle approche de fusion multimodale, appelée approche hybride afin de bénéficier de deux techniques de fusion précoce et tardive.

Dans un second lieu, nous traitons les techniques avancées d'agrégation, en proposant une approche de méta-apprentissage multimodal basée sur le classifieur « *Stacking* » pour le diagnostic automatique du diabète dans un contexte de données déséquilibrées.

Les résultats expérimentaux confirment que l'intégration de la multimodale dans les différentes contributions est capables de générer des inférences plus solides, ou même de nouvelles idées, ce qui serait impossible dans un système monomodal, qui nous permettent de poursuivre dans cette voie de recherche.

Mots-Clés: Classification multimodale, réseaux de neurones convolutionnels (CNN), combinaison de classifieurs, méta-classification, données déséquilibrées.

Abstract

Multimodal classification has become a very important domain of research during the last few years, especially with regard to machine learning and artificial intelligence, thanks to its honorable results achieved through various applications having a practical interest, enabling to integrate and to combine all these modalities of various natures in a coherent set, in a globalizing experience.

The present thesis focuses particularly on the impact analysis as well as on the use of multimodal learning as applied to numerous sources of information on the one hand (textual data and those issued from medical imaging) and, on the other hand, according to the nature of the treated databases (balanced or not) by adopting advanced techniques of machine learning, especially deep learning, as well as ensemble learning approaches with the aim to generate robust systems for decision support.

In a first place, a multimodal approach relying on the convolutional neural networks (*CNN*) and the *SVM* classifier for classification/verification of authors' texts depending on their writing style using "*Word Embedding*"; we additionally suggest, within the medical imaging context, a novel method entitled multimodal data and feature cooperation based classifier fusion for glaucoma diagnosis through the application of a new multimodal fusion approach, referred to as a hybrid approach so as to take advantage of two early and late fusion techniques.

In a second place, we deal with advanced aggregation strategies, by proposing a multimodal meta-learning approach based on the "*Stacking*" classifier as a means to automatically diagnose diabetes within a context of unbalanced data.

Experimental results confirm that integration of multimodal aspect in different contributions is able to generate more robust conclusions, or even new ideas, which would not be possible in a monomodal system, which enables us to pursue our research in this direction.

Keywords: Multimodal classification, convolutional neural networks (CNN), combination of classifiers, meta-classification, unbalanced data.

Table des Matières

Liste des Figures

Liste des Tableaux

Liste des Abréviations

IA	Intelligence Artificiel
SVM	Séparateurs à Vaste Marge
RNA	Réseaux de Neurones Artificiels
PMC	Perceptron multicouche
DL	Deep Learning
ReLU	L'unité linéaire rectifiée
RNN	Réseau de neurones récurrents
LSTM	La mémoire à court terme à long terme
CNNs	Réseaux de Neurones Convolutionnels
R-CNN	réseaux neuronaux convolutionnels récurrents
CONV	La couche de convolution
POOL	La couche de pooling
FC	La couche entièrement connectée
LOSS	La couche de perte
DNN	Deep Neural Networks
DBN	Deep Belief Network
MCS	système multi-classifieurs
K-ppv	k-plus proches voisins
NB	Naîf de Bayes
GLCM	Matrice de co-occurrence des niveaux de gris (Gray Level Co-occurrence Matrix)
GLRLM	Matrice de longueurs de plages des niveaux de gris (Gray Level Run Length Matrix)
HOS	Higher Order Spectra

ADL	Analyse Discriminante Linéaire
ACP	analyse en composantes principales
SFS	sélection séquentielle avant
SBS	sélection séquentielle arrière
EKF	filtre de Kalman étendu
SVF	fusion vectorielle d'état
RIM-ONE	Base de données rétinienne pour l'évaluation du nerf optique (Retinal Image Database for Optic Nerve Evaluation)
TALN	Traitement Automatique du Langage Naturel
VM	vote à la majorité
VMP	vote à la majorité pondérée
VPMP	vote pondéré meilleur-pire
WE	Word Embeddings ou plongement lexical
CBOW	sac de mots continu
DMNB	Discriminative Multinomial Naive Bayes
MSE	Erreur quadratique moyenne
AUC	L'air sous la courbe ROC (Area Under ROC Curve)
ROC	Fonction d'efficacité du récepteur (Receiver Operating Characteristic)
RVB	Rouge, vert, bleu
SMOTE	Synthetic Minority Over-sampling Technique
RESAMPLE	Technique de sur-échantillonnage
CCM	Coefficient de Corrélation Matthews
SOM	Carte auto-organisatrice
SBC	Selective Bayesian classifier
MRMR	Minimum Redundancy Maximum Relevance

Introduction Générale

Ces derniers temps, divers modèles ont été proposés pour booster les performances des systèmes d'aide à la décision, qui se basent principalement sur des techniques d'intelligence artificielle. L'incertitude et l'imprécision constituent les enjeux les plus complexes en matière d'aide à la décision, notamment en ce qui concerne le diagnostic médical. Par ailleurs, la classification reposant sur l'apprentissage automatique s'est imposée comme une solution incontournable dans le cadre des systèmes informatiques d'aide à la décision, en particulier dans un contexte multimodal.

En effet, de nos jours, de nombreuses applications du monde réel requièrent la mise en place d'un traitement multimodal des données. En outre, les informations recueillies auprès du monde réel se composent de manière inhérente de données de différentes modalités.

L'approche multimodale en matière de classification vise à construire des modèles en mesure de traiter et de mettre en relation des informations provenant de modalités multiples. Cela permet aux systèmes la collecte et l'analyse d'une série de données hétérogènes/disjointes issues de divers capteurs et entrées de données en un modèle unique en vue d'une prise de décision plus fiable; par le biais également de l'apport d'informations complémentaires.

Ce type d'apprentissage fait désormais l'objet d'un sujet de recherche brûlant qui est adopté dans de nombreuses applications présentant un intérêt pratique. La notion de « *multimodalité* » désigne le recours à plusieurs modalités permettant d'accomplir la même tâche ou afin de parvenir au même objectif.

Le terme « *modalité* » fait référence à une forme concrète spécifique d'un mode de communication, à savoir le mode visuel, le mode sonore, le mode gestuel, le mode textuel, etc. À titre d'exemple, le bruit, la musique et la parole représentent des modalités du mode sonore.

Motivation de la Thèse

En raison de la richesse des caractéristiques des phénomènes naturels, il est rare qu'une seule modalité fournisse une connaissance complète du phénomène d'intérêt. Théoriquement, un système informatique multimodal correspond à un système capable d'intégrer plusieurs modalités, même s'il intègre un seul mode.

L'approche multimodale repose essentiellement sur la fusion d'informations distinctes provenant de diverses sources afin de trouver la meilleure architecture combinant ces dernières dans le but de fiabiliser la prise de décision. De plus, en apportant des informations complémentaires, l'approche multimodale peut non seulement offrir de meilleures performances, mais également plus de robustesse pour diverses tâches de classification.

La fusion multimodale consiste à utiliser des algorithmes permettant de combiner des informations provenant de plusieurs modalités. L'objectif de cette fusion est de parvenir à de meilleures performances des tâches que celles des approches à une seule modalité. Les jeux de données multimodaux sont susceptibles de contenir différents types de données, tels que du texte, des images, des vidéos, des audios, des articles, des actualités, des blogs et des documents XML. Le défi de la fusion multimodale réside alors dans le fait de savoir comment combiner de manière efficace ces données issues de différentes sources et natures.

Pareillement, l'apprentissage profond «*deep learning*», un ensemble de méthodes avancées d'intelligence artificielle, contribue de manière importante à l'efficacité des systèmes de décision conçus. En effet, l'apprentissage profond et plus particulièrement les réseaux de neurones convolutionnels (*CNNs*) rendent pratiquement possible l'extraction à plusieurs niveaux de caractéristiques et de représentations de haut niveau à partir de données qui visent à éviter le recours à des caractéristiques artisanales fastidieuses grâce à l'apprentissage de bout en bout sur des données brutes.

Une des perspectives des techniques de l'apprentissage profond est le remplacement de certains travaux, encore relativement laborieux, par des modèles algorithmiques d'apprentissage supervisé, non supervisé (c'est-à-dire ne nécessitant pas de connaissances spécifiques quant au problème étudié) ou encore par des techniques d'extraction hiérarchique des caractéristiques. Cette approche a démontré sa capacité de fournir des résultats impressionnants pour de nombreux domaines de recherche, notamment la reconnaissance d'images, le diagnostic médical ainsi que le traitement automatique du langage naturel.

Portée de la Thèse et Problématiques

Le cœur de notre travail tout au long de la présente thèse consiste à examiner la question de savoir si les performances des systèmes d'aide à la décision dans le contexte de l'imagerie médicale ainsi que dans celui des données textuelles reposant sur une seule modalité sont susceptibles d'être améliorées en intégrant des informations apprises et complémentaires issues principalement de différentes modalités, en utilisant des paradigmes avancés de l'apprentissage automatique essentiellement l'apprentissage profond.

L'objectif principal de cette recherche sera d'étudier l'efficacité du recours aux techniques avancées d'apprentissage automatique ainsi qu'aux méthodes de fusion de prédicteurs en vue notamment de l'analyse des systèmes d'aide à la décision basés sur des données multimodales, et ce avec les objectifs spécifiques mentionnés ci-dessous:

- Explorer toutes les approches multimodales actuelles.
- Concevoir divers schémas de fusion correspondant à différents niveaux de fusion au moyen des méthodes avancées d'apprentissage automatique.
- Déterminer la structure la plus appropriée pour l'apprentissage profond.
- Développer et évaluer une approche multi-classifiers pour la vérification des auteurs basée sur le prolongement lexical (*word embedding*) et l'apprentissage profond.
- Élaborer un système de classification multimodal en vue du diagnostic du glaucome reposant sur l'apprentissage profond.
- Développer et mettre au point une approche de méta-classification dans le cadre de l'apprentissage déséquilibré.

Cette recherche pose également une question fondamentale: comment déterminer le schéma de fusion le plus performant et dans quelle mesure l'application de techniques avancées d'agrégation et d'apprentissage profond peut permettre d'améliorer les performances?

En effet, l'un des défis auxquels les systèmes de traitement des données médicales et contextuelles sont confrontés est celui de la masse considérable de données à traiter ainsi que de l'hétérogénéité de ces dernières.

Contributions

Au niveau méthodologique, l'analyse multimodale présente des concepts, des méthodes ainsi qu'un cadre permettant la collecte et l'analyse des interactions localisées dans une perspective englobant différentes ressources tout en bénéficiant de leur utilisation simultanée en vue de résoudre un problème donné.

Au cours des deux dernières décennies, il y a eu une évolution significative des machines informatiques ainsi qu'un progrès important des paradigmes de l'intelligence artificielle, que ce soit au niveau des techniques de prétraitement des données relatives aux images et textes ou à celui des architectures et des algorithmes innovants dans le cadre de l'apprentissage automatique. Cette avancée nous incite à relever de nouveaux défis relatifs à la manière de coopérer et de traiter les données multimodales pour une valeur ajoutée dans l'analyse, l'interprétation et le suivi des systèmes d'aide à la décision.

La présente thèse propose une nouvelle démarche qui s'appuie sur des architectures de fusion multimodale parallèle pour la classification des données médicales et textuelles afin de tirer parti de la complémentarité/compatibilité de trois phases, à savoir: modalités, descripteurs représentatifs des données originales (tels que la matrice GLCM, les moments centraux et les moments Hu) ainsi que la combinaison de classifieurs (au niveau de la prédiction ou des modèles innovants de décision). En ce qui concerne les modèles de décision innovants, nous visons à utiliser:

- L'apprentissage profond, en particulier les réseaux de neurones convolutionnels (*CNNs*), un classifieur incontournable d'une part, pour l'extraction automatique de la matrice des primitives et d'autre part comme une solution idéale pour la classification de la modalité *image*.
- Plongement de mots ou plongement lexical (*Word2vec_Word Embeddings*) pour la modélisation du langage et l'apprentissage des descripteurs dans le cadre du traitement du langage naturel (dans le cas de modalité *textuelle*).

La particularité de ce travail consiste en ce que chacune des technologies adoptées au cours de chaque phase s'adapte en fonction de la nature des données traitées. Cette thèse apporte donc les principales contributions originales suivantes:

- La première contribution porte sur une approche multimodale d'aide à la décision qui s'applique à deux types d'informations telles que les données textuelles et celles de l'imagerie médicale en adoptant les réseaux de neurones convolutionnels (*CNNs*) et le classifieur machine à vecteur de support ou *Support Vector Machine* (*SVM*).

D'un coté, nous proposons une nouvelle méthode pour la tâche de vérification d'auteurs basée sur trois architectures différentes (*CNN*, *RCNN*, et *SVM*) à l'aide des plongements de mots (*word2vec*), tout en appliquant le principe de la fusion tardive qui consiste à traiter chaque modalité séparément lors de l'étape de classification et à fusionner ensuite les résultats des scores de décision obtenus pour chaque modalité en utilisant des méthodes d'agrégation.

D'un autre coté, nous suggérons une nouvelle approche de fusion multimodale, appelée *approche hybride*, destinée au diagnostic du glaucome en utilisant la complémentarité pouvant exister entre les modèles (*CNN* et *SVM*), et ce en combinant d'une part les différentes modalités avant l'apprentissage, et d'autre part l'utilisation des classifieurs séparés pour chaque combinaison de modalités dans le but de bénéficier aussi bien des aspects de fusion précoce que tardive dans un contexte multimodal.

- La seconde contribution consiste en la proposition d'une nouvelle approche de méta-classification multimodale se basant sur le classifieur *Stacking* pour le diagnostic automatique de la maladie du diabète dans un contexte de données déséquilibrées. Une phase préliminaire est menée en appliquant une technique de ré-échantillonnage améliorée, appelée *IRESAMPLE+*, dont le but est de résoudre/surmonter le problème du déséquilibre des données en vue d'améliorer les performances du système de classification.

Organisation de la thèse

La présente thèse se répartit de la manière suivante :

- **Chapitre 01 :** ***L'Intelligence Artificielle : vers des Techniques Avancées d'Apprentissage Automatique***

Dans ce chapitre, nous survolons des définitions de l'intelligence artificielle. Nous discutons ensuite des concepts généraux de l'apprentissage profond tout en nous concentrant sur la nature des différentes couches qui le composent.

- **Chapitre 02 :** ***Classification ensembliste multimodale***

Ce chapitre introduit en premier lieu la notion de *multimodalité* ainsi que le formalisme de la classification supervisée dans le cadre de la classification d'ensemble multimodale. De plus, les principales architectures de combinaison (approches séquentielles, parallèles et hybrides) sont décrites, suivis par la présentation des méthodes classiques proposées pour la combinaison parallèle de classifieurs.

- **Chapitre 03 :** ***Analyse de l'Impact de la Multimodalité pour l'Aide à la Décision***

Ce chapitre détaille notre contribution dans cette thèse consistant en la proposition d'une nouvelle approche multimodale d'aide à la décision appliquée à de nombreuses sources/natures d'informations (telles que les données textuelles et celles issues de l'imagerie médicale) au moyen de techniques avancées de l'apprentissage automatique, en particulier l'apprentissage profond ainsi que les approches ensemblistes.

- **Chapitre 04 :** ***Apprentissage Multimodal Déséquilibré dans le cadre du Diagnostic Précoce du Diabète par le biais d'une Technique de Ré-échantillonnage Améliorée***

Ce chapitre aborde la deuxième contribution de cette thèse, qui propose une nouvelle approche de méta-classification multimodale reposant sur le classifieur *Stacking* en vue du diagnostic automatique du diabète en contexte d'apprentissage déséquilibré, en optant pour une méthode de ré-échantillonnage (*IRESAMPLE+*) améliorée.

Enfin, une conclusion générale ainsi que les perspectives envisagées pour la poursuite de ce travail achèvent cette thèse.

CHAPITRE 01

L'INTELLIGENCE ARTIFICIELLE : VERS DES TECHNIQUES AVANCÉES D'APPRENTISSAGE AUTOMATIQUE

CHAPITRE 01. L'Intelligence Artificielle : vers Des Techniques Avancées d'Apprentissage Automatique

1 Introduction

Les années 1950 ont vu naître l'ambition de créer des systèmes ou des machines capables de simuler l'intelligence humaine. En d'autres termes, l'Intelligence Artificielle (IA) vise à générer des systèmes informatiques pouvant comprendre, percevoir et souvent prendre des décisions, et qui ont également la capacité de simplifier voire de remplacer l'intervention humaine.

Depuis plus d'un demi-siècle, la recherche dans le domaine de l'IA s'est considérablement développée. Ce domaine de recherche a de plus en plus attiré l'attention de nombreuses disciplines et chercheurs d'horizons différents et est devenu un sujet largement interdisciplinaire. Avec le développement technologique, l'intelligence artificielle s'est rapidement répandue et développée, et elle a favorisé le développement d'autres disciplines.

Ces dernières années, les ordinateurs modernes ont fait de grands progrès tant sur le plan matériel que logiciel, et la recherche sur l'IA a également joué un rôle important. Bien que l'intelligence artificielle soit confrontée à de nombreux défis dans le processus de développement, les défis coexistent toujours avec des opportunités afin de développer des systèmes d'aide à la décision et d'améliorer la vie humaine.

L'intelligence artificielle présente de nombreux avantages :

- Grâce à sa précision, le principal avantage de l'intégration de l'intelligence artificielle est de réduire le risque d'erreur (moins d'erreurs). De plus, l'IA peut initier une discussion collaborative avec les humains pour les aider à prendre des décisions (les décisions peuvent être prises très rapidement).
- L'intelligence artificielle fournira des pistes, des indicateurs, des données statistiques et des possibilités qui simplifieront le processus de prise de décision, ce qui réduira le risque d'erreur.

- L'intelligence artificielle fournit des outils infatigables. Contrairement aux humains, les machines ne nécessitent pas de pauses ou de rafraîchissements fréquents. Elles sont programmées pour de longues heures et peuvent fonctionner sans interruption.

L'intelligence artificielle est devenue un sujet populaire ces dernières années, en grande partie grâce aux progrès récents de l'apprentissage automatique et en particulier aux réalisations obtenues grâce à l'apprentissage profond « *deep learning* ». Ce chapitre présente une définition de l'intelligence artificielle ainsi qu'une description des techniques avancées d'apprentissage automatique, en mettant l'accent sur l'apprentissage profond « *réseaux de neurones convolutionnels* » et ses différentes couches.

2 Comprendre l'Intelligence Artificielle (IA)

Reconnaissance d'images, reconnaissance vocale, géolocalisation, réponse automatique aux e-mails, proposition d'achat ciblant les goûts des consommateurs, etc. L'intelligence artificielle est déjà à l'œuvre dans notre quotidien, et sa place ne cesse de croître, de s'affirmer à chaque seconde.

Le terme d'Intelligence Artificielle (IA, ou AI en anglais pour Artificial Intelligence) provient du concept de John McCarthy lors du projet de recherche estival de Dartmouth en 1956 [1]. « Artificiel » signifie créé par des êtres humains, et la capacité d'être « Intelligent » signifie la capacité du système à stocker des connaissances et à accomplir des actions. Il est composé d'indicateurs multidimensionnels tels que le raisonnement, la mémoire, l'émotion et l'expression.

L'intelligence artificielle, comme de nombreuses disciplines émergentes, n'a pas encore été définie avec précision. Une définition générale est presque impossible, car l'intelligence semble être un mélange de traitement et d'expression de l'information. L'intelligence artificielle, telle que définie dans le dictionnaire, est un système informatique qui simule les activités intelligentes des êtres humains. Autrement dit, l'IA constitue une vaste branche de l'informatique qui s'intéresse à la construction des machines intelligentes capables d'effectuer des tâches qui nécessitent généralement une intelligence humaine telles que la prise de décision, la détection d'objets, la résolution de problèmes complexes, etc. Un système peut être considéré comme intelligent lorsqu'il a appris à exécuter une tâche liée au processus auquel il a été affecté sans aucune intervention humaine et avec une grande précision.

L'intelligence artificielle est généralement divisée en trois catégories principales:

- Intelligence Artificielle Étroite, également connue sous le nom « d'IA faible », c'est l'étape de l'intelligence artificielle impliquant des machines qui ne peuvent effectuer qu'un ensemble étroitement défini de tâches spécifiques. À ce stade, la machine ne possède aucune capacité de réflexion, elle exécute simplement un ensemble de fonctions prédéfinies. En d'autres termes, une IA faible égale ou surpasse les capacités humaines sur une tâche bien spécifique.
- Intelligence Artificielle Générale, également appelée «IA forte», est une machine dotée d'une intelligence générale et, tout comme un être humain, elle peut appliquer cette intelligence pour résoudre n'importe quel problème. Autrement dit, les machines auront la capacité de penser et de prendre des décisions comme les humains.
- Super Intelligence Artificielle, c'est le stade de l'intelligence artificielle où la capacité des ordinateurs dépassera les êtres humains, c'est-à-dire que les machines sont beaucoup plus intelligentes que les meilleurs cerveaux humains dans quasiment tous les domaines.

L'intelligence artificielle peut être utilisée pour résoudre des problèmes du monde réel en mettant en œuvre les paradigmes suivants: l'apprentissage automatique, l'apprentissage profond et le traitement du langage naturel. En effet, l'apprentissage profond est la technique la plus prometteuse utilisée récemment en intelligence artificielle.

3 Apprentissage Automatique

L'apprentissage automatique est un sous-domaine majeur de l'intelligence artificielle qui vise à étudier, développer, comprendre et évaluer des algorithmes et des techniques qui permettent aux ordinateurs / machines d'apprendre. Cet apprentissage, qui se fait toujours, est basé sur une sorte d'expérience illustrée par / reposant sur des données d'entrée ou des instructions connues. L'apprentissage automatique consiste à faire en sorte que les ordinateurs modifient ou adaptent leurs actions afin qu'elles deviennent plus précises.

Définition : On dit qu'un programme informatique apprend de l'expérience E à l'égard d'une certaine classe de tâches T et de mesure de performance P, si son rendement (sa performance) aux tâches en T, tel que mesuré par P, s'améliore avec l'expérience E [2].

Par conséquent, en général, l'apprentissage consiste à améliorer les performances futures en utilisant l'expérience passée, en réduisant autant que possible l'intervention ou l'assistance

humaine. En fait, le paradigme de l'apprentissage automatique peut être considéré comme une « programmation par démonstration », où l'approche met l'accent sur l'utilisation d'exemples concrets plutôt que sur la description d'une procédure abstraite. Les tâches d'apprentissage automatique sont généralement classées en trois grandes catégories différentes, selon la nature du problème rencontré [3]:

- *L'apprentissage supervisé :* l'apprentissage de l'algorithme est guidé à travers les entrées et leurs sorties souhaitées fournies par un «enseignant» (les données sont étiquetées). L'objectif est de construire une règle qui mappe les entrées dans leurs sorties. On parle de classification si les étiquettes sont discrètes, ou de régression si elles sont continues.

- *L'apprentissage non supervisé :* l'entrée donnée n'est pas étiquetée et le but de l'algorithme est d'inférer une fonction pour décrire la structure ou le modèle caché dans l'entrée.

- *L'apprentissage par renforcement :* les entrées sont un ensemble de rétroactions provenant d'un environnement dynamique auquel l'algorithme est confronté. Le but est d'atteindre un objectif prédéterminé.

Un nouveau domaine de recherche en apprentissage automatique est apparu récemment connu sous le nom d'apprentissage par transfert qui vise à réutiliser à nouveau un modèle développé pour une tâche particulière, et comme point de départ pour un modèle sur une deuxième tâche. Il s'agit d'une approche populaire en apprentissage profond où des modèles préformés sont utilisés comme point de départ pour diverses tâches d'IA (telles que la vision par ordinateur et le traitement automatique du langage naturel).

Définition : L'apprentissage par transfert est l'amélioration de l'apprentissage dans une nouvelle tâche par le transfert des connaissances (caractéristiques, poids, etc.) d'une tâche connexe qui a déjà été apprise et même s'attaquer aux problèmes qui ont moins de données pour la nouvelle tâche [4].

L'apprentissage profond peut être mis en œuvre en tant que technique supervisée ou non supervisée. Le problème posé dans cette thèse recommande une approche supervisée, donc seule cette branche de l'apprentissage automatique sera traitée.

3.1 L'apprentissage supervisé

La figure 1.1 suivante illustre le processus général d'apprentissage automatique:

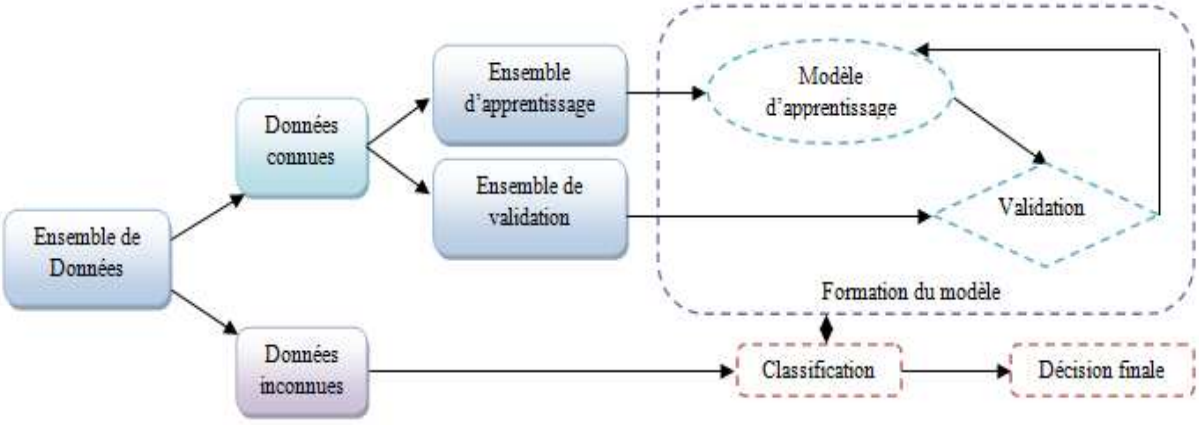

Figure 1.1 : Organigramme de l'apprentissage supervisé

Comme décrit précédemment, l'apprentissage supervisé est une approche d'apprentissage automatique qui nécessite un ensemble de données connues. Cet ensemble comprend des entrées et des sorties appropriées pour l'algorithme utilisé. À partir de cet ensemble d'exemples, le programme est guidé pour décrire un modèle capable de prédire/classifier la sortie correcte. À ce stade, le modèle de prédiction/classification doit être validé avec un autre ensemble de données connues indépendant de l'ensemble d'apprentissage/formation ; ce n'est que lorsque la phase de validation est satisfaisante que l'algorithme peut être considéré comme fiable pour une utilisation sur des données inconnues. Par conséquent, compte tenu d'un problème supervisé et du type de données, les étapes d'apprentissage sont les suivantes:

- *Sélection de l'algorithme :* La première étape consiste à choisir l'algorithme supervisé à utiliser. Chaque méthode a des points forts et des points faibles différents. Le choix dépend du problème particulier, du type et de la quantité de données disponibles. Certains de ces algorithmes sont: Séparateurs à Vaste Marge (SVM), Réseaux de Neurones Artificiels (RNA), et l'apprentissage profond (deep learning). Dans ce travail, l'accent sera mis sur l'apprentissage profond, en particulier sur les réseaux de neurones convolutionnels (CNNs).

- *Formation :* La phase d'apprentissage ou de formation est probablement la plus importante, car les performances finales dépendent du modèle prédictif construit.

- Un ensemble de données connues est sélectionné; il doit être aussi représentatif que possible du problème. L'utilisation d'un ensemble de données pas assez général peut entraîner un sur-apprentissage et de mauvaises performances. Cet ensemble, l'ensemble d'apprentissage, doit fournir une sortie (étiquette) pour chaque entrée répertoriée.
- L'algorithme est entraîné avec l'ensemble de données sélectionné. L'objectif de cette phase est d'essayer de construire un modèle capable de s'adapter aux données fournies, c'est-à-dire de prédire le mieux possible la sortie correcte pour chaque entrée fournie.

- *Validation :* La phase de validation est importante pour tester les performances obtenues par le modèle de prédiction construit lors de la phase précédente.

  - Un autre ensemble de données connues, appelé *ensemble de test*, est préparé. L'ensemble de données doit fournir, en tant qu'ensemble d'apprentissage, des entrées et des sorties fiables pour chaque exemple. Une propriété importante de cet ensemble est qu'il doit être aussi indépendant que possible de celui de l'apprentissage.

  - L'algorithme préalablement formé est utilisé ici pour prédire les données d'entrée de l'ensemble de test. Seules les entrées sont utilisées et les sorties sont prédites par l'algorithme et stockées. La différence fondamentale par rapport à l'étape d'apprentissage est que, dans celle-ci, les étiquettes de sortie ne sont pas utilisées pour améliorer les capacités de prédiction du modèle, mais uniquement pour évaluer ses performances.

  - Les sorties prévues sont validées à l'aide des sorties connues. Les performances sont donc évaluées et analysées. Si elles sont satisfaisantes, il est possible de passer à l'étape finale, sinon l'algorithme ou la phase d'apprentissage doivent être revus avec des précautions ou des paramètres différents.

- *Déploiement du modèle :* une fois l'algorithme formé et validé, il est possible de l'utiliser comme un système automatique pour résoudre le problème d'origine sur les nouvelles données.

3.2 Séparateurs à vaste marge (SVM)

La machine à vecteurs de support ou séparateur à vaste marge (en anglais Support Vector Machine, SVM) est un modèle d'apprentissage supervisé qui a été initialement introduit par *Vapnik* en 1992 [5]. SVM est l'un des algorithmes les plus largement utilisés dans l'apprentissage automatique moderne.

Fondamentalement, l'idée dans SVM est d'adapter une ligne (hyperplan optimal) entre deux classes différentes dans un ensemble de données multidimensionnelles de sorte que la distance entre les niveaux marginaux soit la plus grande taille possible et qu'il n'y ait pas un seul point de données entre les marginaux. Cela a été fait dans la figure 1.2.

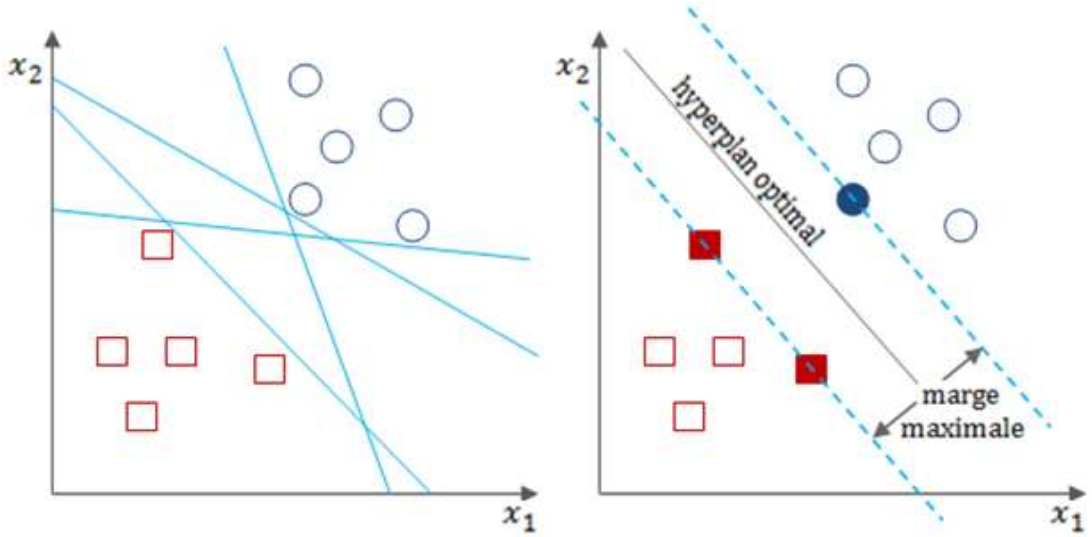

Figure 1.2 : Classification d'un ensemble de données bidimensionnel utilisant SVM [5]

3.3 Apprentissage profond

L'apprentissage profond « *deep learning* » est un ensemble relativement nouveau de techniques d'apprentissage automatique pour les réseaux de neurones multicouches [6]. Il fournit plusieurs algorithmes qui peuvent former des types complexes de réseaux de neurones. Nous pouvons parler d'apprentissage profond lorsque le réseau de neurones a plus de deux couches cachées [6].

L'apprentissage profond est essentiellement une extension des réseaux de neurones artificiels. Il est donc important de comprendre cette technique avant de passer à autre chose.

3.3.1 Réseaux de neurones artificiels (RNA)

RNA est un paradigme de traitement de l'information principalement inspiré par les systèmes nerveux biologiques. Il est composé d'un nombre élevé d'unités de traitement, appelées neurones, travaillant à l'unisson pour résoudre une tâche spécifique. Le processus d'apprentissage en RNA implique, comme dans un système biologique, les ajustements des connexions entre les unités de traitement [2, 6].

3.3.2 Perceptron

L'un des algorithmes des réseaux de neurones artificiels les plus simples est le Perceptron, introduit dans sa version la plus simple par *Frank Rosenblatt* en 1958 [7], comme algorithme d'apprentissage automatique supervisé pour la classification binaire. Il est partiellement inspiré par le neurone biologique et émule en quelque sorte son comportement. C'est la raison pour laquelle ces systèmes sont appelés réseaux de neurones artificiels. La figure 1.3 illustre un neurone biologique.

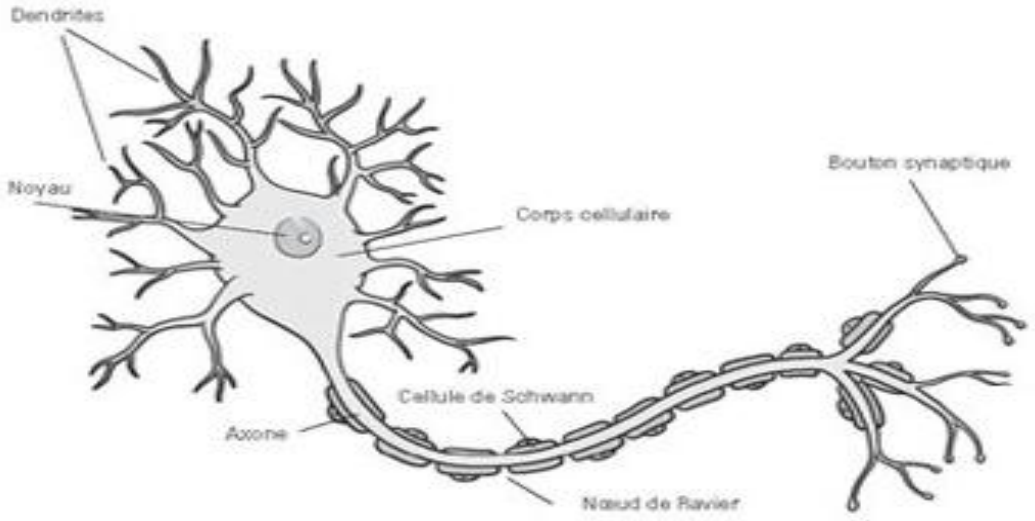

Figure 1.3 : Neurone biologique [7]

Le perceptron original peut être considéré comme un RNA composé d'un neurone artificiel. Il est donc très similaire au concept de neurone artificiel.

Définition: Le neurone artificiel est une fonction mathématique conçue comme un modèle de neurones biologiques.

La figure 1.4 illustre un schéma de Perceptron.

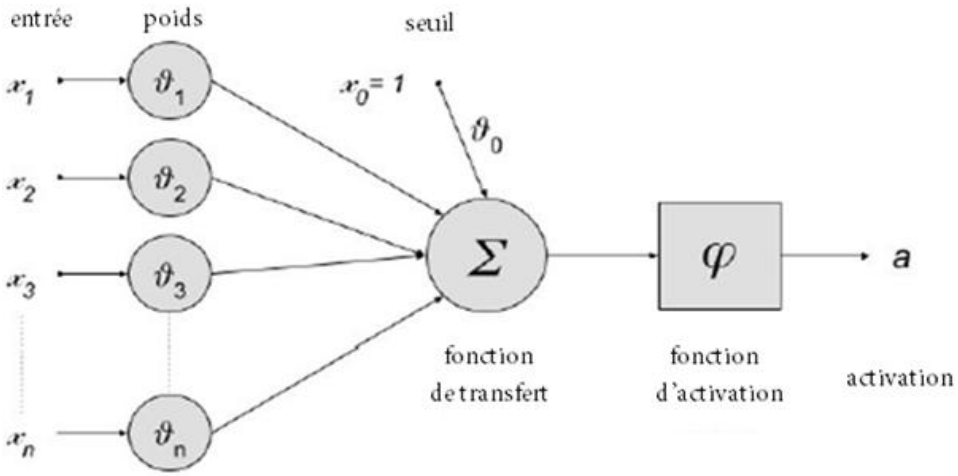

Figure 1.4 : Schéma de perceptron monocouche [2]

Un perceptron prend un vecteur d'entrées en valeur réelle, calcule une combinaison linéaire de ces entrées, puis produit 1 si le résultat est supérieur à un certain seuil et 0 sinon. Plus précisément, étant donné les entrées x_1 à x_n, la sortie $a(x_1, \dots, x_n)$ calculée par le perceptron est:

$$a(x_1, \dots, x_n) = \begin{cases} 1 & si \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n > 0 \\ 0 & ailleurs \end{cases} \tag{1.1}$$

Où chaque θ_i est une constante de valeur réelle, ou poids, qui détermine la contribution de l'entrée xi à la sortie du perceptron. On note que la quantité $-\theta_0$ est un seuil que la combinaison pondérée des entrées $\theta_1 x1 + \theta_1 x2 + \cdots + \theta_n xn$ doit dépasser pour que le perceptron produise a. Pour simplifier la notation, nous imaginons une entrée constante supplémentaire $x_0 = 1$, nous permettant d'écrire l'inégalité ci-dessus comme $\sum_{i=0}^{n} \theta_i x_i > 0$ ou sous forme vectorielle comme $\theta^T . x > 0$. La fonction d'activation d'origine φ mappe la sortie à 1 ou 0. Cette fonction est appelée la *fonction de pas* (voir la figure 1.5):

$$\varphi(x) = step(\theta^T . x) \tag{1.2}$$

Où

$$sgn(z) = \begin{cases} 1 & si \quad z > 0 \\ 0 & ailleurs \end{cases} \tag{1.3}$$

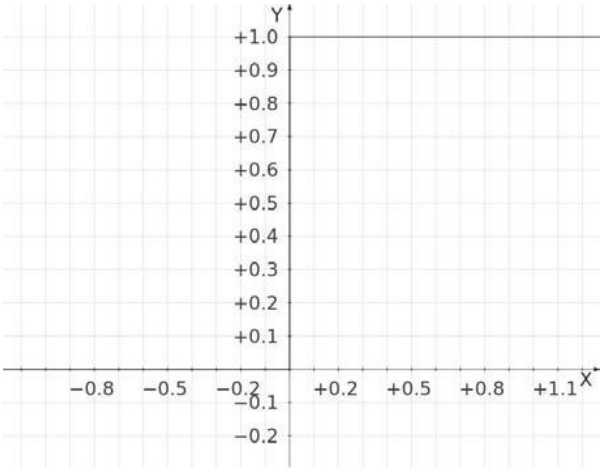

Figure 1.5: fonction de pas

L'apprentissage d'un perceptron implique de choisir des valeurs pour les poids $\theta_0, \ldots, \theta_n$. Par conséquent, l'espace H des hypothèses candidates considérées dans l'apprentissage du perceptron est l'ensemble de tous les vecteurs de poids possibles en valeur réelle $H = \{\theta | \theta \in \mathrm{R}^{\{n+1\}}\}$.

3.3.3 Perceptron multicouche (*multilayer perceptron MLP*)

Les concepts derrière le perceptron sont ceux du réseau neuronal moderne, puis d'apprentissage profond. Une évolution du perceptron est un réseau multicouche avec des couches *cachées* entre l'entrée et la sortie. La figure 1.6 illustre un réseau à 3 couches entièrement connectées (fully-connected).

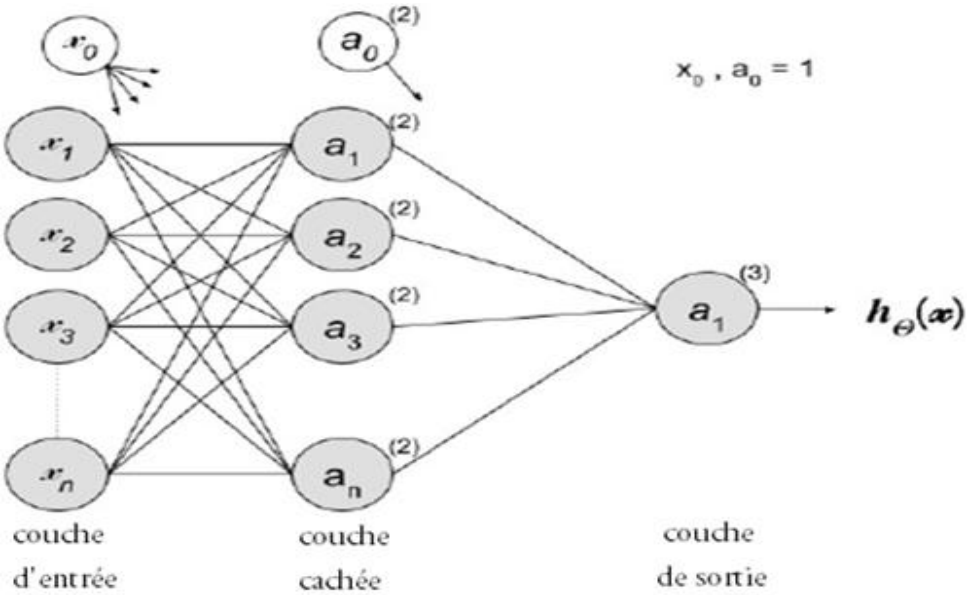

Figure 1.6 : Réseau de neurones à 3 couches entièrement connectées [2]

La première couche est composée de valeurs d'entrée. Pour chaque activation $a_i^{(2)}$ dans la deuxième couche, un vecteur indépendant de poids θ_i^1 est utilisé. Il est donc possible d'écrire :

$$a_i^{(2)} = \theta_i^{(1)}.\ x \tag{1.4}$$

Où θ_i^1 est la ligne i de la matrice θ^1 qui fait correspondre la couche 1 à la couche 2. La fonction d'activation est utilisée pour chaque sortie donnant [2, 7]:

$$a_i^{(2)} = \varphi(\theta_{i0}^{(1)}x_0 + \theta_{i1}^{(1)}x_1 + \theta_{i2}^{(1)}x_2 + \theta_{i3}^{(1)}x_3 + \theta_{i4}^{(1)}x_4) \tag{1.5}$$

Ou, de manière compacte:

$$a^2 = \varphi(\theta^1\ .\ x) \tag{1.6}$$

Chaque activation $a_i^{(2)}$ est ensuite mise en correspondance avec $a_i^{(3)}$ à travers une seconde matrice de poids $\theta^{(2)}$. De plus, il est possible de mapper l'entrée x directement sur la sortie $a_i^{(3)}$en utilisant la notation $a_i^{(3)}= h_\vartheta(x)$ où $\vartheta = \{\theta^{(1)}, \theta^{(2)}\}$.

3.3.4 Fonction d'activation

La fonction d'activation est très importante pour les réseaux profonds [6]. Les fonctions d'activation sont des équations mathématiques qui déterminent la sortie d'un réseau neuronal [2]. La fonction est attachée à chaque neurone du réseau et détermine s'il doit être activé ou non, en fonction de si l'entrée de chaque neurone est pertinente (selon une règle ou un seuil) pour la prédiction du modèle. Les fonctions d'activation permettent également de normaliser la sortie de chaque neurone dans une plage comprise entre 1 et 0 ou entre -1 et 1.

Un autre aspect des fonctions d'activation est qu'elles doivent être efficaces en termes de calcul car elles sont calculées sur des milliers, voire des millions de neurones pour chaque échantillon de données [2]. Les réseaux de neurones modernes utilisent une technique appelée *rétropropagation* pour entraîner le modèle, qui impose une contrainte de calcul accrue sur la fonction d'activation et sa fonction dérivée [2, 6].

Les fonctions d'activation peuvent être fondamentalement divisées en 3 types:

✓ *Fonction de pas binaire (Binary Step Function):* une fonction de pas binaire est une fonction d'activation basée sur un seuil. Si la valeur d'entrée est supérieure ou inférieure à un certain seuil, le neurone est activé et envoie exactement le même signal à la couche suivante

(voir la figure 1.5). Le problème avec une fonction de pas est qu'elle n'autorise pas les sorties à valeurs multiples.

✓ *Fonction d'activation linéaire :* une fonction linéaire est une fonction qui est sur ou presque sur une ligne droite, comme illustré dans la figure suivante:

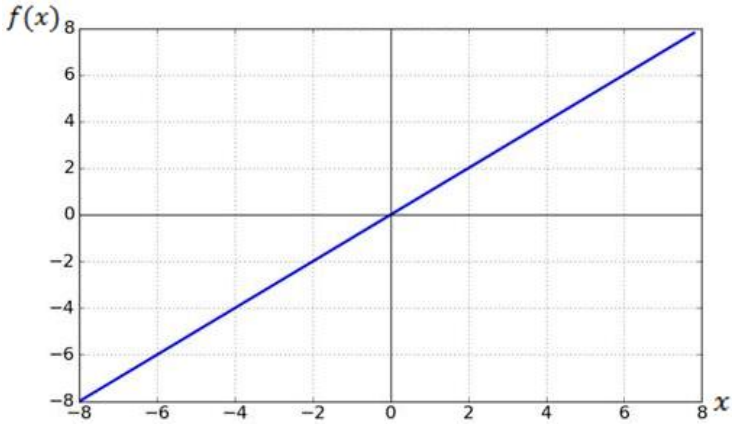

Figure 1.7 : fonction linéaire ou d'identité $\boldsymbol{f(x) = x}$

Elle prend les entrées, multipliées par les poids de chaque neurone, et crée un signal de sortie proportionnel à l'entrée. Dans un sens, une fonction linéaire est meilleure qu'une fonction de pas car elle permet des sorties multiples, pas seulement pas des sorties binaires (oui et non).

✓ *Fonctions d'activation non linéaires :* une fonction non linéaire est une fonction qui n'est pas sur une ligne droite, comme illustré ci-dessous:

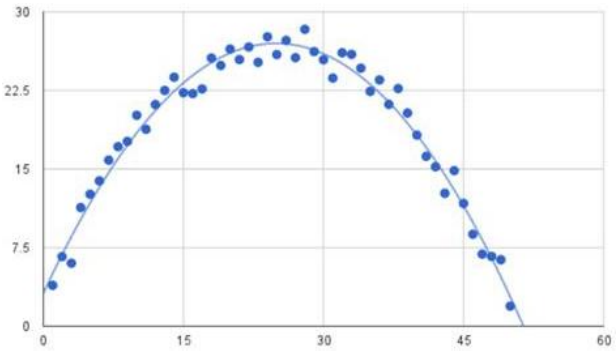

Figure 1.8 : fonction non linéaire

Les fonctions d'activation non linéaires sont les fonctions d'activation les plus utilisées. Elles permettent aux modèles de réseaux de neurones de créer des mappages complexes entre les entrées et les sorties du réseau, qui sont essentiels pour l'apprentissage et la modélisation de données complexes. Les fonctions d'activation non linéaires sont principalement divisées en fonction de leur plage ou de leurs courbes :

- *Fonction d'activation sigmoïde ou logistique :* une fonction sigmoïde est une fonction mathématique ayant une courbe en forme de « S » caractéristique ou une courbe sigmoïde. Un exemple courant de la fonction sigmoïde est la fonction logistique représentée dans la figure 1.9 et définie par la formule:

$$Sig: f(x) = \frac{1}{1 + \exp(-x)} \tag{1.7}$$

Cette fonction permet de propager davantage d'informations initiales et ceci est utile dans le cas de réseaux profonds, et elle est particulièrement utilisée pour les modèles où nous devons prédire la probabilité en tant que sortie.

La fonction *Softmax* est une fonction d'activation logistique plus généralisée qui est utilisée pour la classification multi-classes.

- *Tanh ou fonction d'activation tangente hyperbolique :* une alternative à la fonction sigmoïde logistique est la tangente hyperbolique, ou fonction *Tanh* (figure 1.9) définie par la formule :

$$Tanh: f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{1.8}$$

Comme la sigmoïde logistique, la fonction *tanh* est également sigmoïdale (en forme de « S »), mais génère des valeurs qui varient de -1 à 1. Ainsi, les entrées fortement négatives du *tanh* correspondront aux sorties négatives. En outre, seules les entrées à valeur nulle sont mappées à des sorties proches de zéro. Ces propriétés rendent le réseau moins susceptible de se bloquer pendant la formation.

- *Fonction d'activation d'unité de rectification linéaire (Rectified Linear Unit (ReLU)) :* ReLU est la fonction d'activation la plus couramment utilisée dans les réseaux de neurones, en particulier dans les réseaux de neurones convolutionnels « CNNs » [8]. *ReLU* signifie unité linéaire rectifiée, elle est définie mathématiquement comme:

$$ReLU: f(x) = \max(0, x) \tag{1.9}$$

ReLU est à moitié rectifié (par le bas). Le résultat de $f(x)$ est nul lorsque x est inférieur à zéro et $f(x)$ est égal à x lorsque x est supérieur ou égal à zéro (voir la figure 1.9).

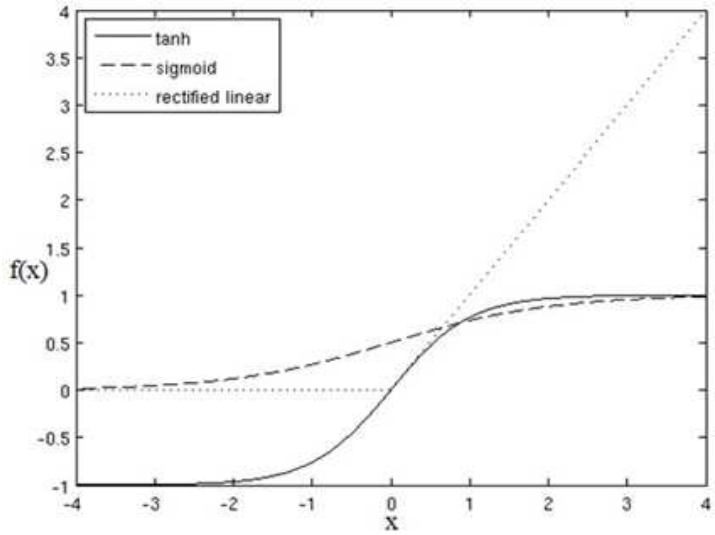

Figure 1.9 : Représentations graphiques de la fonction Tanh, Sigmoïde et ReLU [8]

3.3.5 Algorithme de rétropropagation

La rétropropagation est une méthode d'apprentissage supervisé dans laquelle le réseau ajuste à plusieurs reprises ses poids en fonction de l'erreur ou de l'écart par rapport à la sortie cible, en réponse aux schémas d'apprentissage [2, 6]. La rétropropagation est une forme abrégée de «propagation en arrière des erreurs». Il s'agit d'une méthode standard et la plus courante pour la formation de réseaux de neurones artificiels (RNAs) et de réseaux de neurones convolutionnels (CNNs) [6]. Cette technique permet de calculer le gradient d'une fonction de perte par rapport à tous les poids du réseau. L'apprentissage se déroule à travers plusieurs époques. À chaque époque, les modèles d'apprentissage sont appliqués à la couche d'entrée et les signaux circulent de la couche d'entrée à la couche de sortie à travers des couches cachées.

L'algorithme de rétropropagation comporte trois étapes, à savoir: la propagation avant (*feedforward*) du modèle d'apprentissage, la rétropropagation de l'erreur et l'ajustement du poids [2, 6]. La première propagation consiste à donner au réseau les données d'apprentissage et de comparer la sortie du réseau avec la sortie cible que le réseau allait essayer de prédire, pour générer l'erreur de chaque nœud du réseau. Dans ce cas, chaque neurone d'entrée reçoit un signal d'entrée et le transmet à chaque neurone caché, qui à son tour calcule la fonction d'activation et le transmet à chaque neurone de sortie, qui calcule la fonction d'activation pour obtenir la sortie du réseau. Pendant la phase d'apprentissage, la sortie du réseau est comparée à la cible et l'erreur est calculée. Le facteur d'erreur obtenu à partir de l'erreur est propagé aux couches cachées pour mettre à jour les poids. Ce processus est répété jusqu'à ce que l'erreur soit minimisée.

- ***Étapes de rétropropagation***

L'algorithme de rétro-propagation peut être résumé en quelques étapes. Étant donné un réseau neuronal multicouche avec activation φ, l'algorithme de rétro-propagation nécessite [2, 6]:

- L'ensemble d'apprentissage C.
- Le taux d'apprentissage σ.
- La fonction d'optimisation qui définit l'erreur E.
- Une condition de terminaison, qui peut être un nombre maximal d'étapes ou un taux minimal de réduction des erreurs.

Une fois les exigences initiales accomplies, l'algorithme peut être implémenté comme suit:

Algorithme 1 ***Rétropropagation***

1. Initialiser tous les poids du réseau en petits nombres aléatoires (par ex., entre: $-.05$ et $.05$).
2. Tant que la condition de résiliation est fausse faire
3. Pour chaque $\{x, y\}$de l'ensemble C faire
    - Propager l'entrée vers l'avant via le réseau
4. Entrer l'instance $x^{(i)}$ dans le réseau et calculer la sortie $a^{(l)}$ de chaque nœud du réseau
    - Propager les erreurs en arrière à travers le réseau
5. Calculer l'erreur $\delta^{(L)} = \nabla_a E \emptyset \varphi' (z^{(L)})$ de la sortie du réseau
6. Pour l dans $\{L \dots 1\}$ faire
7. Calculer l'erreur pour chacune des l couches cachées
    - $\delta_i^{(l)} = \left[\left(\theta^{(l)}\right)^T \cdot \delta^{(l+1)}\right] \emptyset \varphi'\left(z^{(l)}\right)$
8. Fin pour
9. Pour l dans $\{L \dots 1\}$ faire
10. Calculer la variation de l'erreur en respectant les poids pour chacune des l couches cachées
    - $\frac{\delta E}{\delta \theta^{(l)}} = \delta^{(l+1)} \cdot \left(a^{(l)}\right)^T$
11. Mettre à jour les couches suivant la descente de gradient stochastique
    - $\theta^{(l)} = \theta^{(l)} - \sigma \cdot \delta^{(l+1)} \cdot \left(a^{(l)}\right)^T$
12. Fin pour
13. Fin pour
14. Fin Tant que.

3.3.6 Réseau de neurones récurrents (RNN)

Les réseaux de neurones récurrents, communément appelés RNNs pour Recurrent Neural Networks, ont été développés pour la première fois au cours des années 1980. Contrairement aux réseaux de neurones ordinaires, les unités d'un RNN ont des connexions - ou boucles - récurrentes qui permettent aux sorties précédentes d'être réinjectées dans le réseau en tant qu'entrée [2,9]. De cette manière, on peut dire que les réseaux de neurones récurrents fonctionnent sur des séquences où chaque étape de la séquence peut représenter une valeur dans le temps ou une position, par exemple un mot ou un caractère dans une phrase.

Nous pouvons imaginer une couche cachée dans un RNN comme un graphe de calcul représentant une chaîne d'événements se déroulant dans le temps, La figure 1.10 illustre un schéma d'un réseau de neurones récurrents à une unité connectant l'entrée et la sortie du réseau.

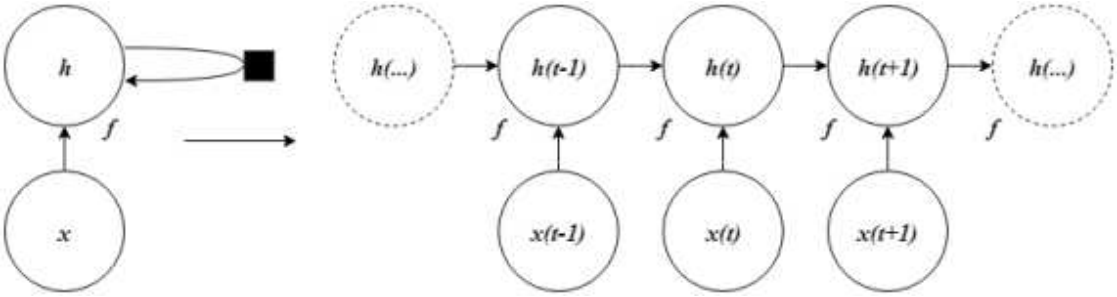

Figure 1.10 : Un modèle d'un réseau de neurones récurrents à une unité, à droite la version « dépliée » de la structure [9]

La sortie à l'étape t (c'est-à-dire de $h^{(t)}$) est envoyée comme entrée à $h^{(t+1)}$ et dépend de la sortie précédente de $h^{(t-1)}$ ainsi que de l'entrée actuelle $x^{(t)}$. Ainsi, les informations sur la sortie aux étapes précédentes sont propagées dans le temps. De cette façon, on peut dire que les réseaux récurrents ont une «mémoire» d'événements survenus plus loin dans le passé.

Un réseau de neurones récurrent est une classe de réseau de neurones artificiels essentiellement utilisé dans le traitement automatique du langage naturel (TALN) et pour l'analyse des séries temporelles dans lesquelles les connexions ne sont pas obligées de passer uniquement de la couche d'entrée à la couche de sortie. Ils peuvent revenir au neurone lui-même, aller à un neurone au même niveau ou aller à un neurone sur une couche précédente. En d'autres termes, les informations peuvent se propager dans les deux directions, y compris des couches profondes aux premières couches. C'est ce qu'on appelle une connexion récurrente [9].

3.3.6.1 RNN à mémoire court-terme et long terme (LSTM)

Bien que les RNNs soient très riches et dynamiques, il est difficile de les former à modéliser « les dépendances à long terme ». Ce dernier a été identifié comme un inconvénient de l'architecture originale RNN dès les années 1990. Cela est dû en partie au problème de gradient qui disparaît et explose. Ces problèmes sont rencontrés en raison de la propagation du gradient à travers de nombreuses couches du réseau récurrent déplié. Des innovations ultérieures ont tenté de résoudre le problème des dépendances à long terme, l'une des plus réussies/efficaces étant la mémoire à court terme et long terme, ou LSTM pour Long Short-Term Memory, proposée pour la première fois en 1997 [10]. La figure 1.11 montre une unité LSTM de base.

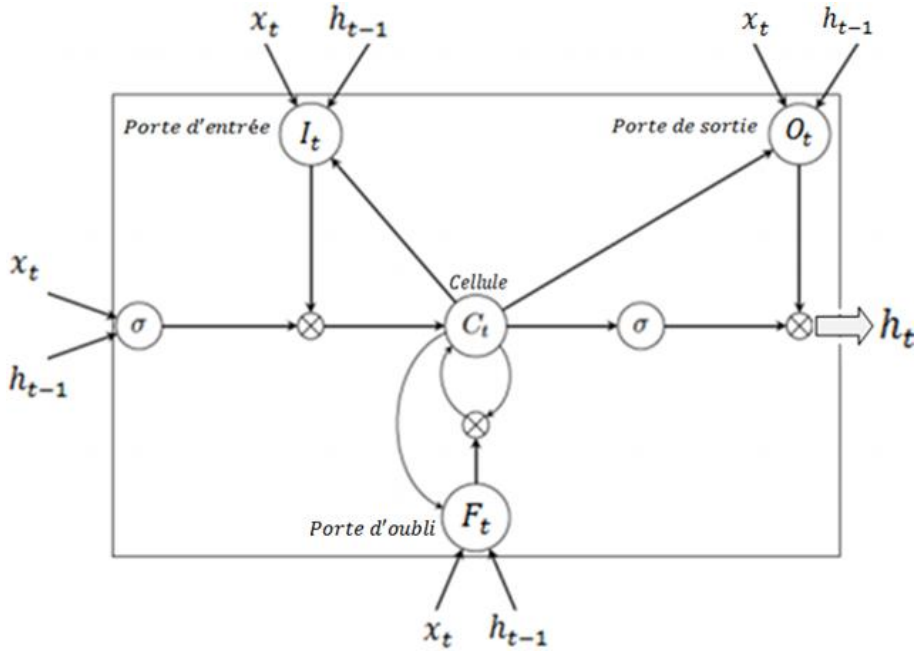

Figure 1.11 : Bloc de mémoire LSTM avec une cellule [10]

Les LSTM sont régis par l'ensemble d'équations suivant:

$$F_t = Sig(W_F x_t + U_F h_{t-1} + b_F) \quad (1.10)$$

$$I_t = Sig(W_I x_t + U_I h_{t-1} + b_I) \quad (1.11)$$

$$O_t = Sig(W_O x_t + U_O h_{t-1} + b_O) \quad (1.12)$$

$$C_t = F_t \cdot C_{t-1} + I_t \cdot Tanh(W_C x_t + U_C h_{t-1} + b_C) \quad (1.13)$$

$$h_t = O_t \cdot Tanh(C_t) \quad (1.14)$$

Dans les équations ci-dessus, x_t est le vecteur d'entrée de la cellule LSTM. F_t, I_t et O_t sont respectivement les vecteurs d'activation de la porte d'oubli, de la porte d'entrée et de la porte

de sortie. L'état de cellule C_t est un composant important dans une cellule LSTM ; il permet à LSTM de transmettre les états historiques traités. Enfin, h_t est le vecteur de sortie de l'unité cellulaire LSTM. Notons que les matrices W, U et le vecteur b sont des poids à apprendre lors de l'apprentissage de cette cellule. Ce b signifie également vecteur de biais.

Les entrées x_t et h_{t-1} sont introduites dans la cellule LSTM, la porte d'oubli F_t détermine d'abord quelles valeurs de l'état de cellule C_{t-1} doivent être rejetées. Ensuite, x_t et h_{t-1} sont passées à la porte d'entrée I_t pour contrôler quand de nouvelles valeurs peuvent circuler dans la mémoire ou les nouvelles valeurs à stocker dans l'état de cellule actuelC_t. Après cela, la porte de sortie O_t calcule quelles informations doivent être produites. Au cours de ce processus, les poids W, U et b sont continuellement mis à jour pour minimiser la fonction de perte.

3.3.6.2 RNN bidirectionnelle à mémoire court-terme et long terme (BLSTM)

Le réseau bidirectionnel à mémoire court-terme et long terme (en anglais Bidirectional Long Short-Term Memory (BLSTM) network), est une variante du modèle LSTM, a été proposé en 1997 [11]. Comme son nom l'indique, le modèle ne prend pas seulement des informations dans les états précédents, il traite les données des états passés et futurs. Par rapport à une couche LSTM standard, un LSTM bidirectionnel ajoute un autre ensemble de cellules LSTM pour traiter les entrées dans une séquence inversée. Les sorties des deux ensembles de cellules sont concaténées et transmises à la couche suivante.

3.3.7 Les réseaux de neurones profonds (Deep Neural Networks (DNN))

Les réseaux DNN sont similaires aux réseaux MLP mais avec plus de couches cachées. Le DNN trouve la manipulation mathématique correcte pour transformer l'entrée en sortie, que ce soit une relation linéaire ou une relation non linéaire. Le réseau se déplace à travers les couches en calculant la probabilité de chaque sortie.

3.3.8 Les réseaux de croyances profondes (Deep Belief Network (DBN))

Un réseau (DBN) est un modèle graphique génératif, ou alternativement une classe de réseau neuronal profond, composé de plusieurs couches de variables latentes «unités cachées», avec des connexions entre les couches, mais pas entre les unités au sein de chaque couche.

Un DBN peut apprendre à reconstruire de manière probabiliste ses entrées dans le contexte d'un apprentissage non-supervisé. Les couches agissent alors comme des détecteurs de

caractéristiques. Après cette étape d'apprentissage, un DBN peut être davantage formé avec supervision pour effectuer la classification.

3.3.9 Les réseaux de neurones convolutionnels (CNNs)

Les réseaux de neurones convolutionnels, désignés par l'acronyme CNNs ou ConvNets, de l'anglais Convolutional Neural Networks sont un type particulier de réseau de neurones utilisés principalement pour classer les images et effectuer la reconnaissance d'objets [6, 8].

L'efficacité des réseaux convolutionnels dans la reconnaissance d'images est l'une des principales raisons pour lesquelles la communauté scientifique a pris conscience de l'efficacité de l'apprentissage profond. Dans un sens, les CNNs sont la raison pour laquelle l'apprentissage profond est célèbre. Le succès d'une architecture convolutionnelle profonde appelée *AlexNet* lors du concours ImageNet 2012[1] a été le coup de feu entendu dans le monde entier. L'émergence et le développement de réseaux de neurones convolutionnels sont à l'origine de progrès majeurs dans le domaine de l'intelligence artificielle, qui ont des applications évidentes pour les voitures autonomes, la robotique, la sécurité et les diagnostics médicaux.

Une autre raison pour laquelle CNN est extrêmement populaire est en raison de son architecture et de sa précision dans diverses tâches de classification difficiles qui nécessitent la compréhension des concepts abstraits dans les images. La meilleure chose qu'il n'est pas nécessaire de passer par l'étape d'extraction artisanale de caractéristiques, le système apprend automatiquement à faire l'extraction de caractéristiques utiles à partir de données brutes [8]. Désormais, les réseaux de neurones convolutionnels peuvent extraire des caractéristiques informatives des images, éliminant ainsi le besoin de méthodes traditionnelles de traitement manuel des images [12].

Le concept de base de CNN est qu'il utilise la convolution de l'image et des filtres pour générer des caractéristiques invariantes qui sont transmises à la couche suivante. Les caractéristiques de la couche suivante sont alambiquées avec différents filtres pour générer des caractéristiques plus invariantes et abstraites, et le processus se poursuit jusqu'à ce que nous obtenions le modèle / la sortie finale qui est invariante aux occlusions.

Les CNNs sont actuellement les modèles les plus efficaces pour la classification d'images / objets. Ils se composent de deux parties distinctes. En entrée, une image est fournie sous la

[1] http://www.image-net.org/challenges/LSVRC/2012/

forme d'une matrice de pixels. Elle a la forme 2D pour représenter une image en niveaux de gris, et de la forme 3D, c'est-à-dire de profondeur 3 pour représenter une image en couleur « Rouge, Vert, Bleu ».

La première phase d'un CNN est la phase convolutive qui est la particularité de ce type de réseaux de neurones. Elle a le rôle d'un extracteur de caractéristiques des images [8, 12]. Une image est passée à travers une succession de filtres, ou *noyaux de convolution*, créant de nouvelles images appelées cartes de convolutions « *feature maps* » (qui sont ensuite normalisées avec une fonction d'activation et/ou redimensionnées). Certains filtres intermédiaires réduisent la résolution de l'image par une opération maximale locale. Au final, les valeurs des dernières cartes de convolutions « *feature maps* » sont mises à plat et concaténées en un seul vecteur de caractéristiques. Ce vecteur définit la sortie de la première phase, et l'entrée de la seconde.

La deuxième phase se compose de couches entièrement connectées qui n'est pas caractéristique d'un CNN, elle se retrouve en fait à la fin de tous les réseaux de neurones utilisés pour la classification. Le rôle de cette phase est de combiner les caractéristiques/valeurs du vecteur (avec plusieurs combinaisons linéaires et fonctions d'activation) pour classer l'image [8, 12]. La sortie est une dernière couche comprenant un neurone par classe. Les valeurs numériques obtenues sont généralement normalisées entre 0 et 1, de somme 1, pour générer une distribution de probabilité sur les classes.

Une architecture de réseau de neurones convolutionnels est composée de plusieurs types de couches de traitement [8]:

- *La couche de convolution (CONV) :* crée une carte des caractéristiques pour prédire les probabilités de classe pour chaque caractéristique en appliquant un filtre qui analyse/balaie l'image entière, quelques pixels à la fois.
- *La couche de pooling (POOL) :* réduit la quantité d'informations générées par la couche *CONV* pour chaque caractéristique et conserve les informations les plus essentielles (le processus des couches *CONV* et de *POOL* se répète généralement plusieurs fois).
- *La couche de correction (ReLU) :* généralement appelée par excès « ReLU » en référence à la fonction d'activation « unité de rectification linéaire », qui remplace toutes les valeurs négatives reçues en entrées par des zéros.

- *La couche entièrement connectée (FC) :* applique des poids sur l'entrée générée par l'analyse des caractéristiques pour prédire une étiquette précise, qui est souvent du type perception.
- *La couche de sortie (FC) ou couche de perte (LOSS) :* génère les probabilités finales pour déterminer une classe pour l'image.

L'architecture d'un CNN est un facteur clé dans la détermination de sa performance et de son efficacité. La façon dont les couches sont structurées, les éléments utilisés dans chaque couche et la façon dont elles sont conçues ont souvent une incidence sur la vitesse et la précision avec lesquelles elles peuvent effectuer diverses tâches.

La figure 1.12 illustre un modèle CNN pour l'aide au diagnostic médical du cancer du sein.

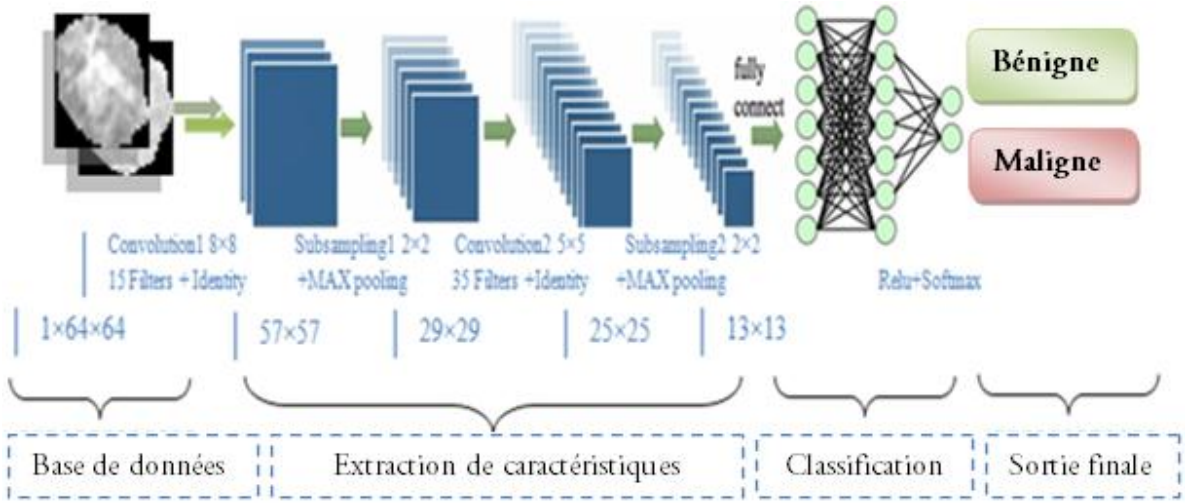

Figure 1.12 : Architecture CNN pour le diagnostic automatique du cancer du sein [12]

3.3.9.1 La couche de convolution

La *couche de convolution (CONV)* représente la principale composante des réseaux CNNs et forme toujours leur première couche. L'objectif de cette couche est de repérer la présence d'un ensemble de caractéristiques dans les images reçues en entrée. À cet égard, un filtrage par convolution est effectué, le principe est de faire *glisser* une fenêtre représentant la caractéristique sur l'image, et de calculer le produit de convolution entre le filtre (*caractéristique*) et chaque portion de l'image balayée [8].

La couche *CONV* reçoit donc plusieurs images en entrée et calcule la convolution de chacune d'entre elles avec chaque filtre. Les filtres correspondent absolument aux caractéristiques que nous voulons trouver dans les images. Pour chaque paire (image à traiter, filtre) une carte des caractéristiques « *feature map* » est obtenue, qui nous indique l'emplacement des

caractéristiques dans l'image, et plus la valeur est élevée, plus l'emplacement correspondant dans l'image est similaire à la caractéristique.

La taille des cartes de caractéristiques à la sortie des couches de convolution dépend des *hyperparamètres*, c'est-à-dire des paramètres qui doivent être définis / régler au préalable. Chaque image possède des dimensions $L \times H \times C$, où L représente sa largeur en pixels, H représente sa hauteur en pixels, et C est le nombre de canaux (1 correspond à une image en noir et blanc et 3 pour une image en couleur « RVB »).
En fait, la couche *CONV* possède / nécessite quatre *hyperparamètres* :

- Le nombre de *noyaux* / filtres « K ».
- La taille des *noyaux* / filtres « T ».
- Le *pas* « $stride$ (S) » avec lequel le noyau glisse sur l'image.
- La marge à 0 « *zero-padding* » P : ajoute une bordure de P pixels de valeur nulle « *zéro* » autour des bords des images d'entrée de la couche.

Pour un volume (image) de taille $L \times H \times C$ à l'entrée, la couche de convolution produit un volume de taille $L_x \times H_x \times C_x$ où:

- $L_x = (L - T + 2P)/S + 1.$
- $H_x = (H - T + 2P)/S + 1.$
- $C_x = K.$

Un réglage commun des *hyperparamètres* est défini comme suit : $T = 3,\ S = 1$ et $P = 1$. En général, le réglage de *zero-padding* à $P = (T - 1)/2$ lorsque le *pas* est $S = 1$ garantit que le volume d'entrée et le volume de sortie auront la même taille dans l'espace.

❖ ***Processus de convolution***

Une couche de convolution L^i (couche i du réseau) est caractérisée par son nombre N de cartes de convolution M_j^i (j $\in\{1, \ldots N\}$), la taille du noyau de convolution « *Kernel* » $K_x \times K_y$ (souvent carré) et le schéma de connexion à la couche précédente L_{i-1}. Chaque carte de convolution M_j^i est le résultat d'une somme de convolution des cartes de la couche précédente M_j^{i-1} par son noyau de convolution respectif. Un biais b_j^i est ensuite ajouté et le résultat est transmis à une fonction de transfert non linéaire $\phi(\ )$. Dans le cas d'une carte complètement connectée aux cartes de couches précédentes, le résultat est alors calculé par [8]:

$$M_j^i = \phi(b_j^i + \sum_{n=1}^{N} M_n^{i\ -\ 1} * K_n^i) \quad (1.15)$$

Où, * est l'opérateur de convolution.

Avec,

$$\phi(x) = 1.7159\ tanh\ (⅔x) \quad (1.16)$$

Le noyau « *kernel* », ou filtre, est une matrice de valeurs, généralement de petites dimensions et souvent carrées, utilisé pour mettre en évidence certaines caractéristiques d'une image donnée. Ce filtre / noyau sera déplacé par pas « S » successifs sur l'ensemble d'images. Pour chaque position du noyau, les valeurs des deux matrices en superposition (noyau et image à traiter) sont multipliées. Chaque valeur ainsi déduite est projetée dans une nouvelle matrice. Cette matrice représente une nouvelle image qui met en évidence les caractéristiques recherchées à travers le filtre / noyau.

Le nombre de filtres détermine le nombre de caractéristiques détectées. Ce nombre est appelé profondeur. Autrement dit, si x filtres sont appliqués à une image donnée, sa valeur de profondeur est x. L'utilisation de filtres est la base d'un réseau *CNN*. En pratique, de nombreux filtres sont testés avec des valeurs différentes pendant la phase d'apprentissage, et les meilleurs filtres sont sélectionnés en fonction du jeu de formation [8].

La figure 1.13 montre un exemple de convolution. La grille bleu ciel représente la carte des caractéristiques d'entrée et, dans cet exemple, une seule carte est représentée pour rendre le dessin facile à comprendre. Un noyau de valeur (zone bleu foncé) glisse sur la carte avec un noyau *« kernel* (K) *»* de taille 3×3 appliqué à une entrée (I) 5×5 en utilisant des pas « *strides* (S) » de 1×1. Les résultats finaux de cette opération sont appelés cartes de caractéristiques de sortie [8].

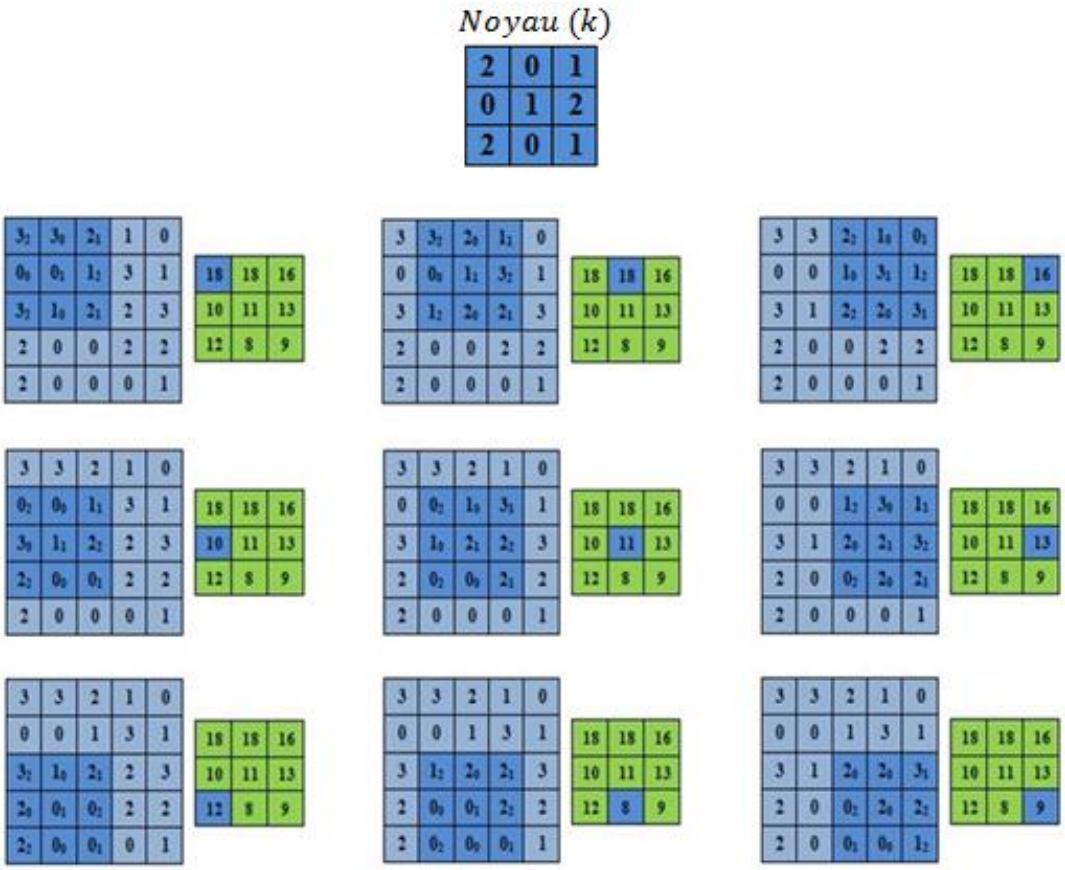

Figure 1.13 : Exemple de calcul des valeurs de sortie d'une convolution [8]

Le schéma de la figure 1.14 montre un exemple de la première étape de convolution.

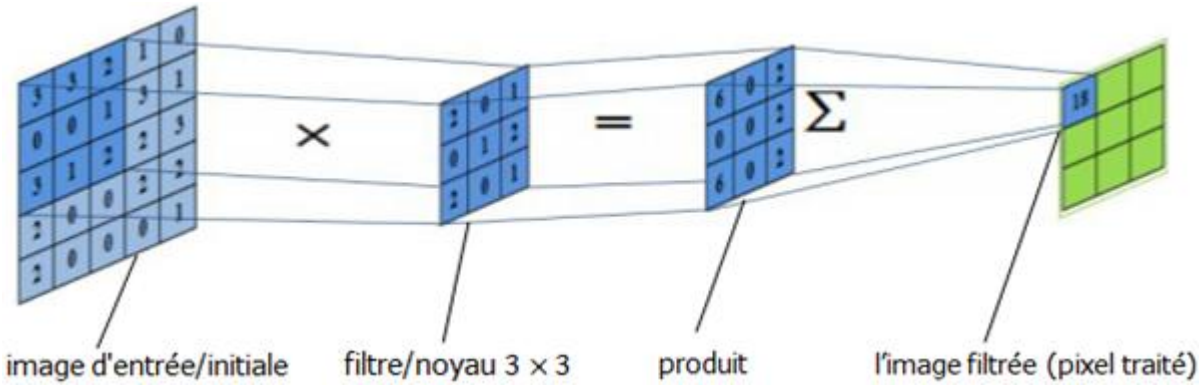

Figure 1.14 : Une étape du calcul de convolution [8]

Le pas « S » définit la façon dont un filtre glisse autour du volume d'entrée et la marge à 0 « *zero-padding* » le remplit de zéros autour de la bordure. Ces pas et remplissage-zéro sont donc utilisés pour contrôler la dimension spatiale du volume de sortie. La figure 1.15 illustre un exemple d'utilisation du « *padding* (P)» et des pas « S »avec une vision plus claire.

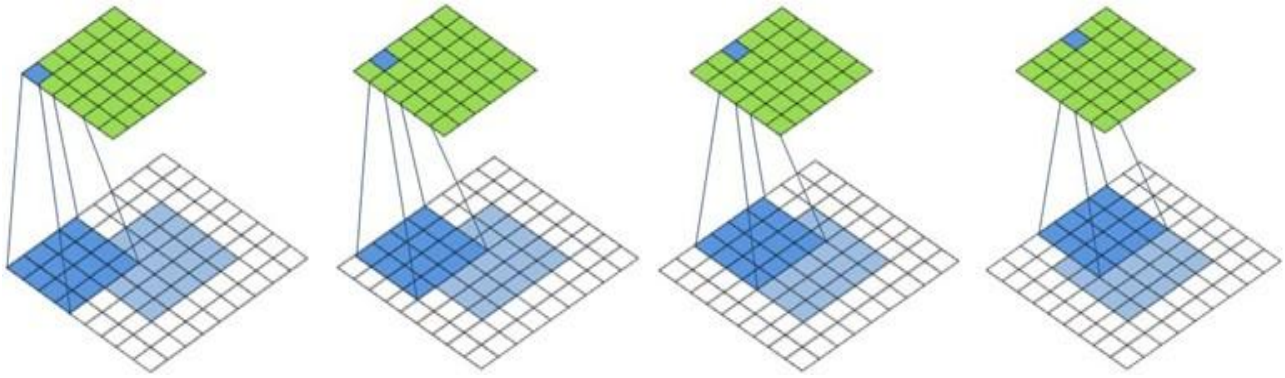

Figure 1.15 : Convolution avec $\boldsymbol{P = 2}$, $\boldsymbol{K = 4}$, $\boldsymbol{I = 5}$ et $\boldsymbol{S = 1}$ [8]

❖ ***Comment initialiser et apprendre les éléments de la matrice de filtrage « features » ?***

Lors de l'apprentissage d'un réseau CNN à partir de zéro, les éléments de filtre des couches sont généralement initialisés à partir d'une distribution aléatoire. Il existe différentes manières d'initialiser le noyau / filtre de la couche de convolution et tout dépend de l'architecture du réseau [8] : des méthodes simples comme l'initialisation *gaussienne*, et des techniques plus avancées telles que l'initialisation *Xavier/Glorot et* l'initialisation *He*.

L'apprentissage à l'aide d'une distribution aléatoire permet d'initialiser les neurones pour qu'ils soient différents et aussi d'apprendre une hiérarchie riche et diversifiée de caractéristiques.

L'apprentissage est la procédure de réglage ou d'ajustement des valeurs des éléments de filtre. De cette façon, les valeurs de filtre sont ajustées pendant l'apprentissage et le système est réputé avoir convergé lorsque la perte (*LOSS*) est minimisée.

La perte est calculée dans la couche finale du réseau en utilisant une méthode de *descente de gradient stochastique* avec une technique d'optimisation avancée comme *Momentum* (algorithme largement utilisé par la communauté d'apprentissage profond) qui permet une meilleure rétropropagation des signaux de gradient [8].

En d'autres termes, les caractéristiques ne sont pas prédéfinies au préalable, mais apprises par le réseau lors de la phase de formation. Les noyaux désignent les poids de la couche de convolution qui sont initialisés puis mis à jour par l'algorithme de *rétropropagation*.
Par conséquent, l'efficacité des réseaux de neurones convolutionnels réside dans la capacité d'identifier les éléments discriminants de l'image, en s'adaptant au problème posé [8, 12].

3.3.9.2 La couche de pooling

La couche de pooling est un concept majeur et un processus essentiel dans un réseau convolutionnel, se situant souvent après la couche de convolution. Elle reçoit en entrée plusieurs cartes de caractéristiques et applique le processus de pooling à chacune d'entre elles. Ce processus consiste à extraire les valeurs essentielles des pixels en réduisant la taille de l'image tout en conservant les caractéristiques pertinentes [8].

La couche de pooling est une forme de sous-échantillonnage de l'image permettant également de diminuer la quantité de paramètres et de calcul dans le réseau, améliorant ainsi l'efficacité du réseau et évitant le sur-apprentissage.

Un autre concept clé de la couche de pooling est de fournir une invariance translationnelle car, en particulier dans les tâches de reconnaissance d'image, la détection de la caractéristique est plus importante par rapport à l'emplacement exact de la caractéristique [8, 12]. Par conséquent, l'opération de pooling vise à préserver les caractéristiques détectées dans une représentation plus petite et le fait, en éliminant les données moins importantes au détriment de la résolution spatiale.

La couche de pooling nécessite un réglage de deux hyperparamètres: la taille des filtres « T » et le pas « S ». Pour un volume (image) de taille $L \times H \times C$ à l'entrée, la couche de pooling génère un volume de taille $L_x \times H_x \times C_x$ où:

- $L_x = (L - T)/S + 1.$
- $H_x = (H - T)/S + 1.$
- $C_x = C.$

Il convient de noter que pour les couches de pooling, il n'est pas courant d'utiliser le concept de *zero-padding*. Un réglage couramment adopté pour les *hyperparamètres* est défini comme suit : $T = 3$ et $S = 2$ et plus communément $T = 2$ et $S = 2$.

Trois approches courantes utilisées dans le processus de pooling qui sont:

- *Average_pooling:* calculer la valeur moyenne pour chaque position représentée par le filtre dans la carte des caractéristiques.
- *Max_pooling :* calculer la valeur maximale pour chaque zone représentée par le filtre dans la carte des caractéristiques.
- *Pooling_stochastique :* similaire à la fonction *Max_pooling*, mais basée sur une méthode probabiliste.

L'approche largement adoptée pour cette couche est le *« Max_pooling »* qui a donné des résultats satisfaisants [8, 12]. L'objectif de cette fonction consiste à redimensionner / diminuer l'image tout en conservant les plus grandes valeurs de pixels. Pour ce faire, un filtre se déplace sur la surface de l'image ou des cartes caractéristiques. Dans chaque région représentée par le filtre, la valeur la plus élevée / validée est extraite, créant ainsi une nouvelle matrice / image de sortie avec seulement les valeurs remarquables de l'image.

La sortie de la couche *Max-pooling* est donnée par la valeur d'activation maximale dans la couche d'entrée pour différentes régions de taille $K_x \times K_y$ qui ne se chevauchent pas. De manière similaire à une couche de convolution, un biais est ajouté et le résultat est transmis à la fonction de transfert $\phi(\ )$ définie ci-dessus [8]. La figure 1.16 illustre un exemple de calcul des valeurs de sortie de la fonction *Max-pooling*.

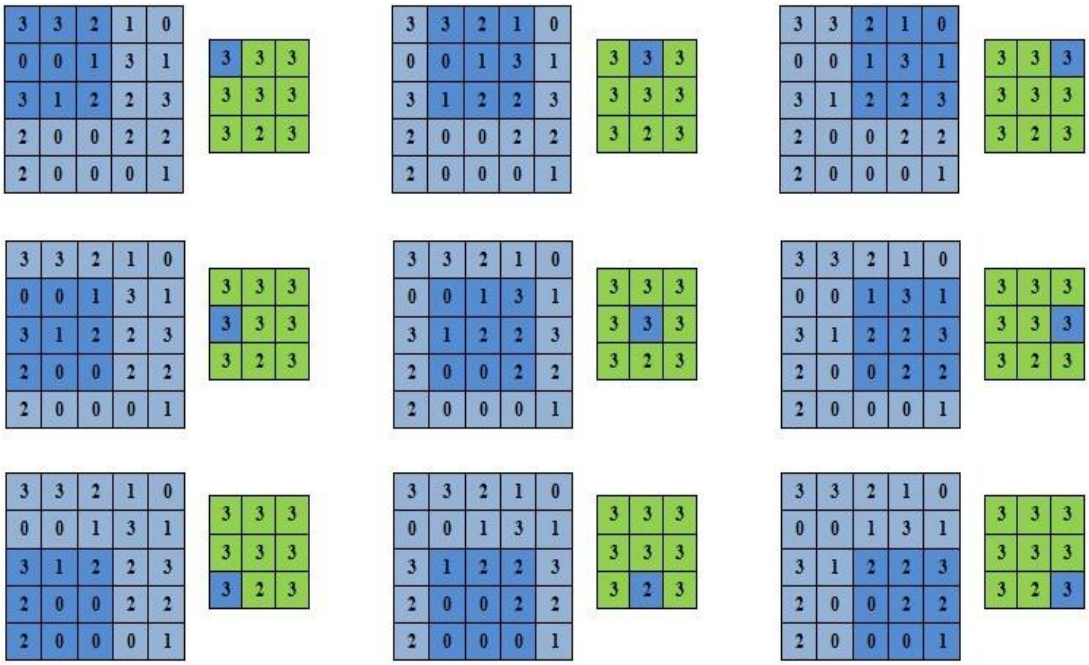

Figure 1.16 : Processus de max_pooling 3×3 sur une entrée 5×5 utilisant des pas de 1×1 [8]

3.3.9.3 La couche de correction (ReLU)

Une couche de correction / rectification dans un réseau de neurones convolutionnels consiste en une fonction d'activation qui prend la carte des caractéristiques générée par la couche de convolution et crée la carte d'activation comme sortie afin d'améliorer l'efficacité du traitement. Indépendamment de la simplicité générale de l'opération, elle joue un rôle important dans les performances du réseau CNN en éliminant les effets d'annulation dans les couches suivantes.

La fonction d'activation est une opération élément par élément sur le volume d'entrée et donc les dimensions de l'entrée et de la sortie sont identiques. La fonction d'activation est généralement mise en œuvre en tant que fonctions *logistiques* « *sigmoïdes* » ou *tangentes hyperboliques*. Cependant, des recherches plus récentes suggèrent que les unités linéaires rectifiées « *ReLU* » sont avantageuses par rapport aux fonctions d'activation traditionnelles, en particulier dans les réseaux de neurones convolutionnels [6, 8, 12].

La fonction *ReLU* transforme les éléments négatifs en zéro, et résout donc le problème d'annulation ainsi que le résultat en un volume d'activation beaucoup plus clairsemé à sa sortie. En outre, *ReLU* consiste uniquement en des opérations simples en termes de calcul (principalement des comparaisons) et donc beaucoup plus efficace à mettre en œuvre dans les réseaux de neurones convolutionnels.

3.3.9.4 La couche entièrement connectée (fully-connected « FC »)

Une fois que les étapes de convolution et de pooling achevées, une carte de caractéristiques regroupées doit être mise en place. Littéralement, nous aplatissons la carte de caractéristiques regroupées dans une colonne afin de l'insérer dans la couche *FC* qui constitue toujours la dernière phase du réseau de neurones convolutionnels pour réaliser la classification.

- **La couche d'entrée *FC* : Aplatissement « *Flattening* »**

L'aplatissement transforme une matrice bidimensionnelle de caractéristiques en un vecteur qui peut être introduit dans un classifieur de la couche *FC*. En d'autres termes, ce processus représente la couche d'entrée de la couche *FC* qui consiste à aplatir toutes les cartes de caractéristiques issues du réseau de convolution dans un seul vecteur [8]. La figure 1.17 montre un exemple d'aplatissement d'une matrice.

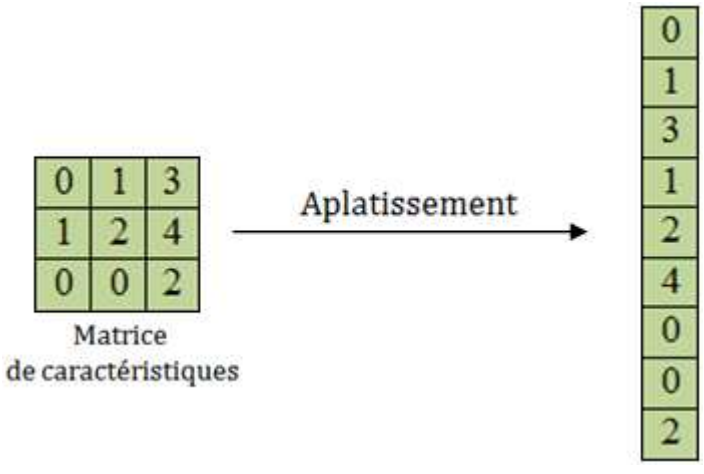

Figure 1.17 : Processus d'aplatissement d'une carte de caractéristiques

Toutes les matrices / cartes de caractéristiques sont ainsi placées bout à bout dans ce même vecteur. Ensuite chacune des valeurs de ce vecteur sera connectée aux neurones de la première couche du réseau *FC* permettant la classification de l'image.

❖ **Principe de la couche FC**

Une couche entièrement connectée (couche FC) est généralement un type de réseau neuronal où tous les neurones des couches adjacentes sont entièrement connectés par paire, mais les neurones de la même couche ne partagent aucune connexion [8]. Cette couche reçoit le vecteur créé lors de l'étape d'aplatissement (la sortie de la couche précédente) en entrée et génère un nouveau vecteur en sortie. Pour ce faire, elle applique une combinaison linéaire puis éventuellement une fonction d'activation aux valeurs reçues en entrée. Une rétropropagation est appliquée à chaque itération de l'apprentissage. Au cours d'une série d'époques, le modèle est capable de distinguer entre les caractéristiques dominantes et certaines caractéristiques de bas niveau dans les images. Autrement dit, la partie entièrement connectée du réseau CNN passe par son propre processus de *rétropropagation* pour déterminer les poids les plus précis. Chaque neurone reçoit des poids qui priorisent l'étiquette la plus appropriée. Enfin, les neurones *votent* sur chacune des étiquettes, et le vainqueur de ce vote est la décision finale de classification.

3.3.9.5 La couche de perte (LOSS)

La couche de perte / erreur représente la couche de sortie *FC* et considérée comme la dernière couche du réseau CNN [8]. Elle définit la manière de mesurer la différence entre la sortie attendue et la sortie réelle lors de la formation du réseau. Il existe de nombreuses fonctions adoptées à des tâches spécifiques : la fonction *Softmax* largement utilisée pour la classification multi-classes qui attribue une catégorie à chaque objet à classer, la fonction *euclidienne* utilisée dans un problème de régression et la fonction *entropie croisée sigmoïde* utilisée pour classer les valeurs de probabilité *P* dans l'intervalle[0,1]. *Un réseau de neurones convolutionnels « CNN » se distingue d'un autre par la manière dont les couches sont structurées, mais également par la façon dont elles sont paramétrées* [6, 8, 12]. Il existe différentes architectures dans le domaine de *réseau convolutionnels* qui ont été essentielles dans la construction d'algorithmes qui alimentent et alimenteront l'intelligence artificielle dans son ensemble dans un avenir prévisible. Les plus courants sont: *LeNet*, *AlexNet*, *VGGNet*, *GoogLeNet*, *ResNet*, *ZFNet*. Le tableau 1 illustre les principales études suggérées dans la littérature utilisant l'apprentissage profond « *DL* » pour diverses tâches de classification.

Tableau 1.1: principales études proposées dans la littérature utilisant l'apprentissage profond

Référence	**Tâche d'application**	**Méthode *DL* utilisée**	**Brève description**
Yang et al. [13]	Diagnostic médical	CNN	Conception d'un système permettant d'identifier simultanément la présence d'un cancer de la prostate dans une image IRM et de localiser les lésions en se basant sur des caractéristiques CNNs et d'un classifieur SVM_RBF.
Oh et al. [14]		CNN + LSTM	Développement d'un système automatisé pour la classification des arythmies en 5 catégories en utilisant des battements ECG de longueur variable en adoptant une combinaison de CNN et LSTM.
Kaya [15]		CNN : *AlexNet* + *VGG16*	Système d'aide à la décision pour la classification des états épileptiques à travers les signaux EEG utilisant CNNs, la technique mRMR (pour la sélection des caractéristiques) et le classifieur K-NN.
Liu et Guo [16]	Classification des textes	CNN + LSTM bidirectionnel (BILSTM)	Proposition d'une nouvelle méthode *AC-BiLSTM* basée sur l'attention avec une couche de convolution afin d'améliorer la compréhension sémantique et la précision de la classification.
Nedjah et al. [17]		CNN + Word Embeddings	Proposition d'un nouveau modèle d'analyse des sentiments utilisant des réseaux CNNs via Word Embeddings en analysant l'impact des hyperparamètres sur les performances du modèle afin d'obtenir un meilleur taux de classification.
Jain et al. [18]		CNN + LSTM	Modélisation d'une nouvelle architecture hybride *SSCL* d'apprentissage profond pour la classification du spam dans les médias sociaux basée sur les réseaux CNNs et LSTM à l'aide de bases de connaissances telles que WordNet et ConceptNet.

4 Conclusion

Dans ce chapitre, nous avons présenté les principaux concepts de l'intelligence artificielle en détaillant certains algorithmes d'apprentissage automatique, en mettant l'accent sur les réseaux de neurones convolutionnels « *CNNs* » et soutenus par des travaux connexes.

Les réseaux *CNNs* sont des algorithmes très puissants qui sont appliqués dans un large éventail de domaines tels que la classification des images et des textes, ainsi que la détection d'objets. La structure hiérarchique et les puissantes capacités d'extraction de caractéristiques font de « *CNN* » un algorithme très robuste pour diverses tâches de classification.

Les performances des architectures convolutionnels profondes « *CNNs* » sont à l'origine de progrès fondamentaux dans les applications actuelles d'intérêt pratique de l'Intelligence Artificielle « *IA* ».

De plus, les algorithmes d'apprentissage automatique visent à utiliser toutes les modalités de données disponibles dans la formation afin d'améliorer les mesures de performance de classification et de fournir une meilleure prise de décision en utilisant des techniques d'apprentissage d'ensemble. En général, l'utilisation du concept de combinaison de classifieurs a pour but de fournir un niveau de performance supérieur à celui de n'importe quel classifieur de base unique. Dans le chapitre suivant, nous développerons la classification ensembliste multimodale.

CHAPITRE 02

CLASSIFICATION ENSEMBLISTE MULTIMODALE

CHAPITRE 02. Classification Ensembliste Multimodale

1 Introduction

La classification d'ensemble est devenue un sujet crucial au cours des dernières décennies, principalement grâce à l'intégration du concept *multimodal*, ce qui a montré un grand potentiel pour améliorer la précision et la fiabilité de diverses tâches de classification présentant un intérêt pratique.

De nos jours, de nombreuses applications du monde réel nécessitent le traitement de données multimodales. En plus, les informations collectées dans la vie réelle sont intrinsèquement constituées de données issues de différentes modalités (en décrivant un concept unique de différentes manières), telles qu'une image Web avec des descriptions de texte narratif vaguement liées, et un article d'actualités avec du texte et des images appariés.

Dans diverses disciplines, des informations sur un phénomène ou un système d'intérêt peuvent être obtenues à partir de différents types de détecteurs, dans multiples conditions, ou dans plusieurs expériences ou sujets, ou d'autres types de sources. Dans ce cas, chaque cadre d'acquisition est désigné comme une « *modalité* » des phénomènes ; celle-ci est associée à un ensemble de données. En raison de la relation complexe et riche entre les modalités des phénomènes multimodaux, une seule modalité ne peut pas décrire suffisamment l'événement d'intérêt. Le fait que plusieurs modalités rendent compte du même événement introduit de nouveaux défis qui dépassent les degrés de liberté associés à l'exploitation de chaque modalité séparément.

Généralement, un système est considéré comme *multimodal* lorsqu'il reçoit des données de deux sources ou plus qui décrivent le même concept différemment et effectue une analyse individuelle sur les deux.

Avec l'abondance de données multimodales disponibles sur internet au cours des deux dernières décennies, les chercheurs ont développé de nombreux algorithmes et modèles de fusion pour intégrer des données issues de sources multiples pour diverses tâches de classification (comme le diagnostic des pathologies et la classification des textes).

L'apprentissage automatique multimodal vise à construire des modèles capables de traiter et de relier l'information multimodale afin d'offrir la possibilité de capter les correspondances entre les modalités et d'acquérir une compréhension profonde du phénomène naturel.

La fusion multimodale est l'un des sujets originaux de l'apprentissage automatique multimodal qui consiste à utiliser des paradigmes d'apprentissage d'ensemble « *ensemble learning* » pour combiner des informations provenant de différentes modalités « *multimodalités* », dont l'objectif est d'obtenir une meilleure performance et plus de robustesse des tâches par rapport aux approches à modalité unique.

Ce domaine de recherche présente des défis uniques sur la façon de combiner/fusionner efficacement les données provenant de différentes sources/natures compte tenu de l'hétérogénéité de celles-ci. Il existe différentes méthodes de fusion des modalités, où la fusion des caractéristiques et la fusion des décisions sont souvent les plus adoptées dans les approches multimodales. Ce chapitre illustre les différentes stratégies de fusion multimodale ainsi qu'une description du concept de *multimodalité*, tout en présentant les différentes notions de base de la combinaison de classifieurs.

2 Multimodalité ?

Le concept de *multimodalité* fait référence à l'utilisation de plusieurs *modalités* pour effectuer la même tâche, quelle que soit la nature des modalités. La modalité est un concept vague qui peut se concrétiser de différentes manières. En sémiotique, la définition de la modalité est la suivante :

> *Une modalité est une façon particulière de coder l'information en vue de la présenter. Elle fait référence à un certain type d'information et/ou au format de représentation dans lequel l'information est stockée* [19].

En général, une modalité fait référence à la manière dont quelque chose se produit ou est vécu. Le mot *modalité* est habituellement associé à des modalités sensorielles qui représentent les principaux canaux de communication et de sensation. En d'autres termes, une modalité est une forme concrète spécifique d'un mode de communication (par exemple, auditif, spatial, gestuel, visuel ou linguistique). Un problème de recherche est donc caractérisé comme *multimodal* lorsqu'il est capable d'intégrer plusieurs de ces modalités (même s'il n'intègre qu'un seul mode).

Une propriété fondamentale de la *multimodalité* est connue sous le terme de « *complémentarité* », en ce sens que chaque modalité apporte à l'ensemble un certain type de valeur ajoutée qui ne peut être déduite ou obtenue à partir d'aucune des autres modalités du système. En termes mathématiques, cette valeur ajoutée est connue sous le nom de « *diversité* ». La diversité permet de réduire le nombre de degrés de liberté dans le système en fournissant des contraintes qui améliorent l'unicité, l'interprétabilité, la robustesse, la prise de décision, la performance et d'autres propriétés souhaitées [20].

3 Fusion multimodale

La fusion multimodale est le processus de combinaison de données/informations provenant de modalités/sources multiples afin d'inférer de nouveaux résultats qui ne peuvent être obtenus par aucune des sources uniques ou d'obtenir des résultats plus efficaces et précis que n'importe laquelle des sources uniques. La figure 2.1 montre un système de fusion multimodale typique.

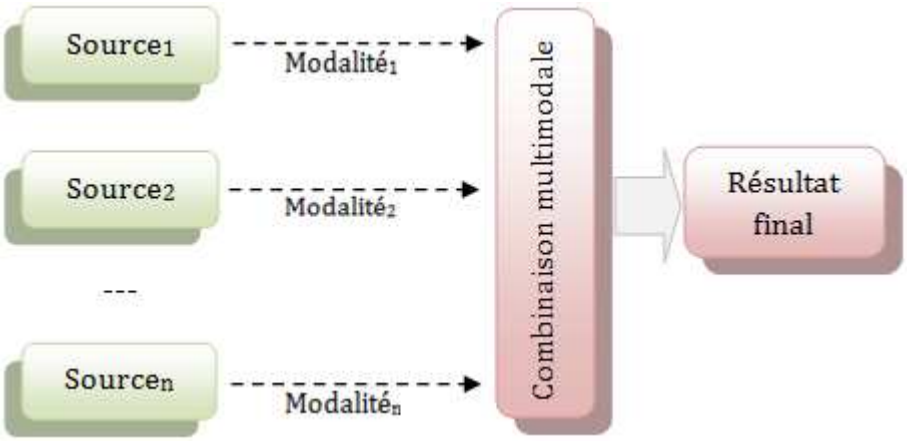

Figure 2.1 : Un système de fusion multimodale typique

3.1 Conception d'un système multimodal

Dans un système de fusion multimodale comme le montre la figure 2.1, la « modalité/source » peut être représentée par l'un des éléments suivants : capteur, caractéristique, classifieur, ressource d'information (ensemble de données, bibliothèque, différentes situations du monde réel, etc.) [21]. Ce qui suit sont des explications de chaque type de source avec des exemples correspondants:

- *Capteurs :* le système de fusion combine les sorties de plusieurs capteurs. Par exemple, combinant les différentes sorties/modalités d'imagerie médicale TC, IRM et PET.
- *Caractéristiques:* Le système de fusion combine des instances de différentes caractéristiques « features ». Par exemple, combinant des descripteurs/caractéristiques de couleur, de texture et de forme pour reconnaître un objet. De plus, les caractéristiques « CNN », « couleur », « forme » et « texture » extraites de la source des images (médicales) résument l'information sous différents aspects, et, par conséquent, ces caractéristiques sont acceptées en tant que des modalités différentes, lorsque cela est nécessaire. En bref, le critère pour être une modalité est d'avoir un aspect différent de la représentation des données et une quantité importante d'informations complémentaires avec d'autres modalités.
- *Classifieurs :* Le système de fusion combine plusieurs classifieurs différents. La création de différents classifieurs peut se faire de plusieurs manières. Duin [22] présente des méthodes de génération de différents classifieurs. Certains d'entre eux incluent l'utilisation d'algorithmes de classification différents, ayant différents choix de paramètres, un ensemble d'apprentissage différent, effectuant différentes initialisations, etc. Par exemple : combinant les résultats d'un classifieur CNN et d'un SVM pour les mêmes données d'entrée, ou combinant des classifieurs de réseaux neuronaux artificiels différemment initialisés.
- *Ressources d'information:* le système de fusion peut combiner les informations de différents ensembles de données / instances / situations. En d'autres termes, le système peut combiner plusieurs sorties d'un seul capteur, différentes instances d'une seule caractéristique, des classifieurs formés avec différents jeux de données (algorithme de classification, paramètres, initialisations, etc.) ou différentes instances d'une seule modalité. Par exemple, combinant deux classifieurs d'arbre de décision qui sont formés sur différents ensembles de données.

La fusion multimodale est l'analyse de plusieurs modalités/ensembles de données de telle sorte que différents ensembles de données peuvent interagir et s'informer mutuellement. La fusion est un processus à plusieurs niveaux et à multiples facettes visant à la détection automatique, l'association, la corrélation, l'estimation et la combinaison d'informations provenant de diverses sources singulières et plurielles en vue d'améliorer la prise de décision.

3.2 Intérêt de la fusion multimodale

Le paradigme bien accepté selon lequel certains processus et phénomènes naturels peuvent s'exprimer sous des formes physiques complètement différentes est la raison d'être de la fusion multimodale de données. Les motivations pour la fusion multimodale sont nombreuses. Il s'agit notamment d'obtenir une image plus unifiée et une vue globale du système, ainsi que d'améliorer la prise de décision. Certaines des raisons pour lesquelles le paradigme multimodal est préférable à une seule modalité/source sont illustrées ci-dessous:

- La fusion de sources complémentaires offre une représentation plus complète du système. La fusion de sources multimodales (coopératives ou concurrentes) réduit l'incertitude et augmente la robustesse [23].
- Un avantage pratique de la fusion est qu'elle réduit les sources non fiables. Ainsi, la fusion permet de réduire la dépendance à l'égard de n'importe laquelle des sources [23].
- Le bruit dans les données détectées entraîne des inefficacités dans la reconnaissance; le fait d'avoir plusieurs sources peut diminuer l'effet du bruit [24].
- Chaque système à source unique a une limite supérieure de performance. Les performances de classification d'un système à modalité unique ne peuvent pas être améliorées en continu en réglant l'extraction des caractéristiques, le classifieur ou certaines autres étapes. De cette façon, aucune source ou méthodologie unique n'est complètement parfaite, une combinaison de ces sources/modalités peut non seulement donner de meilleurs résultats mais également plus de fiabilité [24, 25].

3.3 Niveaux de fusion multimodale

La fusion multimodale peut être décrite comme la combinaison d'informations, souvent imparfaites et hétérogènes, en vue de se procurer une information globale utile, plus complète, de meilleure qualité, et permettant d'améliorer la prise de décision [23, 24]. La fusion d'informations dans un système multimodal peut prendre différentes formes selon le moment où elle est effectuée. La figure 2.2 montre qu'il existe cinq possibilités de fusion multimodale dans la littérature [20, 23, 25, 26, 27, 28, 29]. Il est possible de fusionner des données (brutes) directement au niveau du capteur/acquisition, de fusionner des instances de différentes caractéristiques ou de fusionner des informations au niveau du score, du rang ou pendant la phase de prise de décision.

Il convient de noter que dans un système de fusion, les informations disponibles diffèrent pour chaque niveau de fusion. Aux niveaux inférieurs, davantage d'informations sont disponibles, mais l'utilisation de ces informations aussi détaillées conduit à un système complexe en termes de calcul ; tandis qu'aux niveaux supérieurs, moins d'informations sont fournies pour le processus de fusion et, par conséquent, elles sont plus faciles à combiner, ce qui se traduit par un système moins complexe. En outre, les niveaux inférieurs procurent plus de gains par fusion que les niveaux supérieurs, en raison des informations utilisables/disponibles. Les sous-sections suivantes développent les différents niveaux concernant la fusion multimodale.

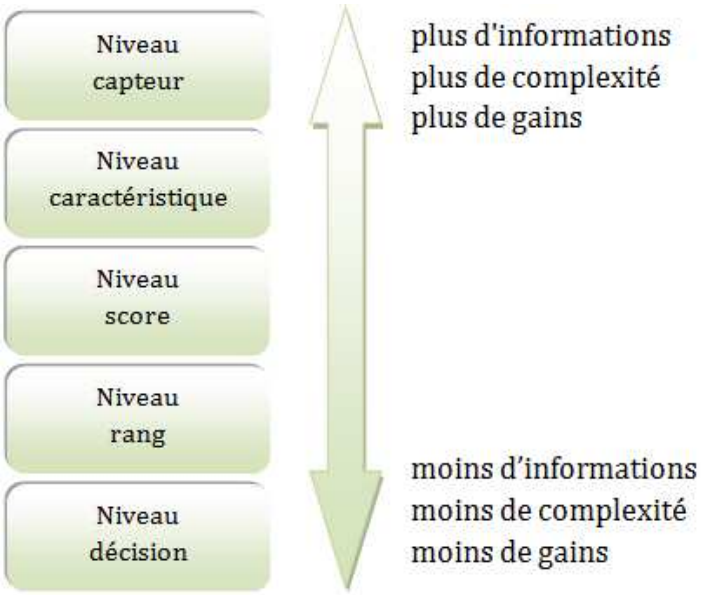

Figure 2.2 : Niveaux de fusion multimodale

3.3.1 Niveau capteur

La fusion au niveau du capteur représente le niveau de fusion ayant la source d'information la plus riche, mais le processus de fusion est le plus complexe (il nécessite une certaine homogénéité entre les données). Ce niveau implique la combinaison de données brutes acquises à partir de plusieurs capteurs, et elle peut être réalisée si les sources multiples représentent des échantillons du même phénomène d'intérêt obtenu à l'aide de capteurs compatibles uniques ou différents [27].

Les exemples de fusion au niveau du capteur comprennent une combinaison de plusieurs capteurs (par ex. fusion d'images avec des capteurs de couleurs RVB et des capteurs NIR) et/ou une combinaison de plusieurs instantanés avec un seul capteur (par ex. fusion d'images

utilisant des images prises à différents moments/changements de pose avec un seul capteur qui pourrait ensuite être soumis à l'extraction de caractéristiques).

3.3.2 Niveau caractéristique

La fusion au niveau des caractéristiques est une sorte de solution qui combine différents vecteurs de caractéristiques, obtenus soit avec différentes multimodalités, ou bien en appliquant différents algorithmes/descripteurs d'extraction de caractéristiques à la même modalité (par ex. caractéristiques préformées « CNNs » et caractéristiques traditionnelles « forme, texture, etc. ») afin de créer un nouvel ensemble de caractéristiques qui devrait être la synthèse des informations les plus pertinentes pour guider au mieux le classifieur lors de la prise de décision finale [26, 27].

Cette stratégie de combinaison est généralement mise en œuvre à travers le processus de *concaténation* des vecteurs de caractéristiques extraits de multiples sources de données. Cela permet de générer un ensemble de vecteurs de grande taille qui est généralement coûteux en calcul. Une procédure de sélection/transformation des caractéristiques peut être adoptée pour obtenir un ensemble minimal de caractéristiques à partir du vecteur de caractéristiques de grande dimension.

3.3.3 Niveau score

La fusion à ce stade est également connu sous le terme de fusion au niveau de mesure/confiance représentant le niveau intermédiaire de la fusion multimodale qui intègre les scores (de correspondance) fournis par plusieurs classifieurs en utilisant une modalité unique/multiple via un schéma d'agrégation (de type mesure) afin d'obtenir un nouveau score qui est ensuite utilisé pour prendre la décision finale [27, 28, 29].

La fusion au niveau du score est généralement préférée dans les systèmes multimodaux car les scores (de correspondance) contiennent suffisamment d'informations à propos du modèle d'entrée. En réalité, ce type de fusion offre le meilleur compromis entre la richesse d'information et la facilité de mise en œuvre.

Habituellement, un processus de normalisation des scores est requis pour mettre à l'échelle les scores résultant (par ex. mesure de distance ou de similarité) de différentes modalités dans la même plage, de sorte qu'aucune modalité unique ne prévaut sur les autres, et que l'importance de chaque modalité individuelle est mise à profit dans la décision finale.

3.3.4 Niveau rang

La fusion au niveau du rang est semblable à celle de la fusion au niveau score, sauf que les informations utilisables pour ce niveau de fusion deviennent moindres et seuls les rangs pour les résultats de la classification sont disponibles. On fait référence à la fusion au niveau des rangs lorsque la sortie de chaque classifieur/matcher (étape de classification) est un sous-ensemble de correspondances possibles triées dans un ordre décroissant de confiance (un rang plus élevé indiquant une bonne correspondance) [22, 29].

Autrement dit, ce type de fusion implique de consolider les multiples rangs associés à une catégorie et de déterminer un nouveau rang qui aiderait à établir la décision finale. L'utilisation de listes de rang en tant que données d'entrée pour la fusion rend le processus de fusion beaucoup plus simple puisque l'utilisation de listes de rang ne nécessite pas un processus de normalisation et que les listes de rang de différentes modalités/classifieurs sont directement comparables. Ainsi, à ce niveau de fusion, il est plus simple de mettre en œuvre un système de fusion que le niveau de score.

3.3.5 Niveau décision

À ce niveau de fusion, chacun des classifieurs représentant les différentes modalités fournit une décision individuelle (c'est-à-dire que les classifieurs sont traités séparément) de la meilleure correspondance possible en fonction des données qui lui sont présentées. Pour les systèmes de classification binaire, il s'agit d'une étiquette de classe ou d'une décision vraie ou fausse qui peut être représentée respectivement par 1 et 0. Après cela, une méthode d'agrégation appropriée est utilisée pour combiner les différentes sorties des classifieurs en vue de générer la décision finale du système multimodal [22, 29].

La fusion d'informations au niveau abstrait ou au niveau décision peut être mise en place s'il existe au moins trois modalités/classifieurs. En d'autres termes, le nombre de modalités doit être impair pour éviter le cas de l'égalité entre les décisions des différents classifieurs.

La fusion au niveau décisionnel représente le niveau le plus élevé et l'approche la plus simple pour implémenter un système multimodal. Cependant, la fusion à ce stade contient moins d'informations disponibles et moins de gains de fusion par rapport à d'autres niveaux de fusion multimodale.

En effet, trois classes principales de systèmes multimodaux peuvent être différenciées en fonction des informations qu'elles combinent, à savoir:

- *Système multi-capteurs*: plusieurs capteurs sont utilisés pour acquérir/représenter la même modalité.
- *Système multi-instances*: plusieurs instances de caractéristiques diverses sont employées pour la même modalité.
- *Système multi-classifieurs*: plusieurs algorithmes d'apprentissage automatique traitent une modalité unique/multiple utilisant un ou de nombreux extracteurs/vecteurs de caractéristiques, qui peuvent intervenir dans l'un des modules de correspondance (score, rang ou décision).

3.4 Méthodologies de fusion multimodale

La fusion multimodale est une méthode populaire pour augmenter les performances de différentes applications d'intérêt pratique en consolidant les informations fournies par plusieurs modalités/classifieurs. En particulier, la stratégie de fusion devrait être capable de tirer pleinement parti des informations recueillies auprès de sources multiples et de mieux décrire la méthodologie envisagée. Un système multimodal mal conçu produira probablement des performances dégradées et une faisabilité réduite.

La méthodologie de fusion définit quel algorithme est utilisé pour combiner l'information multimodale. La littérature comprend de nombreuses méthodes proposées et expérimentées pour la fusion multimodale. Cependant, il est difficile de prédire si une combinaison est meilleure/supérieure, donc aucune préférence claire d'une technique de combinaison n'est compromise [30].

Les méthodes basées sur des filtres de *Kalman* tels que le filtre de *Kalman étendu* (EKF), le filtre de *Kalman non parfumé* (UKF), la *fusion vectorielle d'état* (SVF), la *fusion de mesure* (MF) et la *fusion de gain* (GF) sont des techniques courantes pour combiner les informations de différentes capteurs « *système multi-capteurs* » [31, 32].

En ce qui concerne les méthodes utilisées au niveau caractéristique, une procédure de *concaténation* des caractéristiques est largement adoptée pour combiner les instances de plusieurs ensembles de caractéristiques « *système multi-instances* » [33], dans le but de générer un seul vecteur de caractéristiques sur lequel l'apprentissage est effectué. Cette procédure peut être suivie d'une technique de sélection (comme la *sélection séquentielle avant* (SFS), la *sélection séquentielle arrière* (SBS), etc.) ou de normalisation/transformation de caractéristiques (comme *analyse en composantes principales* (ACP), *analyse discriminante*

linéaire (ADL), etc.) sur le vecteur de caractéristiques résultant afin de réduire sa dimensionnalité [33].

Chibelushi et al. [34] ont suggéré un schéma visant à combiner les caractéristiques associées à la voix (audio) et à la forme des lèvres (vidéo) d'un individu dans un système d'identification. Les transformations ACP et ADL sont utilisées pour réduire la dimensionnalité de l'ensemble de caractéristiques concaténées. Du et al. [35] ont proposé une architecture de réseau générale appelée mécanisme de connexion de caractéristiques sélectives (SFCM) en vue de concaténer les caractéristiques profondes des différentes couches *CNN* d'une manière simple et efficace appliquée à plusieurs tâches difficiles de vision par ordinateur, y compris la classification des images, la détection des textes et des scènes.

Les algorithmes employés pour combiner l'information multimodale auprès des systèmes multi-classifieurs sont généralement répartis en deux catégories : les méthodes paramétriques (basées sur des techniques d'agrégation linéaire et de vote) et les méthodes non-paramétriques (basées sur l'apprentissage/la classification) [22, 25, 36, 37]. La fusion au niveau du score, du rang et lors de la phase de décision « *système multi-classifieurs* » est considérée comme l'approche largement adoptée pour les études multimodales utilisant les techniques d'apprentissage d'ensemble « *combinaison de classifieurs* » en raison de sa simplicité de mise en œuvre et de ses performances supérieurs pour diverses tâches de classification [22, 25, 29].

3.5 Classification d'ensemble

Récemment, la classification d'ensemble, souvent fait référence à la *combinaison de classifieurs*, a été proposée comme une voie de recherche cruciale rendant la reconnaissance plus fiable en utilisant la complémentarité qui peut exister entre les classifieurs. La classification d'ensemble est un ensemble de classifieurs dont les décisions individuelles sont combinées pour classer de nouveaux exemples. Une combinaison de classifieurs est souvent beaucoup plus précise que les classifieurs individuels qui les composent. Sur ce point, la littérature abonde de travaux présentant des méthodes de combinaison qui se différencient aussi bien par le type d'informations apportées par chaque classifieur que par leurs capacités d'apprentissage et d'adaptation. Avant de développer le concept de la combinaison de classifieurs, quatre points sont utiles pour déterminer la perspective:

1. Type d'informations de sortie du classifieur à combiner.
2. La structure de combinaison.
3. La nature des classifieurs à combiner.
4. La règle de combinaison utilisée pour combiner les décisions des classifieurs.

3.5.1 Motivation

Dans le domaine de l'apprentissage automatique, l'aspect le plus important de chaque algorithme de classification est sa capacité à généraliser, ce qui signifie sa capacité à étiqueter correctement un échantillon de test (nouvel échantillon) en utilisant les connaissances acquises à partir des données d'apprentissage. En 1974, Kanal [38] a mentionné pour des tâches de classification: "*No single model exists for all pattern recognition problems and no single technique is applicable to all problems. Rather what we have is a bag of tools and a bag of problems*", et d'après le théorème de *no free lunch theorem* énoncé par *Wolpert* [39], il est claire qu'il n'y a pas d'algorithme de classification universellement optimal et, par conséquent, il existe une vaste littérature sur les techniques d'apprentissage supervisé, et la combinaison de classifieurs découle de l'imitation de la nature humaine, qui a tendance à rechercher plusieurs opinions avant de prendre une décision cruciale.

La combinaison de classifieurs a fait l'objet d'une attention croissante ces dernières années, de nombreuses études théoriques et empiriques démontrant l'efficacité des ensembles sur un seul classifieur dans différentes circonstances [40-42]. Conformément à Dietterich [43], il existe plusieurs raisons théoriques et pratiques pour lesquelles un système multi-classifieurs est préférable à un seul classifieur. Ces raisons peuvent être classées en trois types de limitations, à savoir: *statistiques*, *représentationnelles* et *computationnelles/de calcul*.

- *Raisons statistiques* : une limitation statistique d'un algorithme d'apprentissage survient lorsque la quantité de données d'apprentissage disponibles est trop petite par rapport à la taille de l'espace d'hypothèse. Sans données suffisantes, l'algorithme d'apprentissage peut trouver plusieurs fonctions d'hypothèse qui donnent la même précision sur les données d'apprentissage mais ne sont pas en mesure de généraliser par elles-mêmes en présence de nouvelles données. En d'autres termes, il existe une grande *variance* dans le processus de sélection de la fonction de décision. Cependant, la création d'un ensemble de classifieurs (à partir de tous ces classifieurs) dénotés par $\mathcal{F}_{app}$ et l'agrégation de leurs décisions par une approche appropriée peuvent réduire le risque de choisir un classifieur médiocre et de sur-apprentissage.

- *Raisons représentationnelles* : Au contraire, dans de nombreuses applications, le modèle appris est trop simple pour approximer la fonction d'hypothèse optimale (la distance « *attendue* » entre la fonction apprise et la fonction optimale est grande). Cela équivaut à avoir un problème de *biais* important. En effectuant une combinaison entre différentes hypothèses, il peut être possible d'élargir l'espace des fonctions représentables et donc de réduire le sous-apprentissage. En d'autres termes, l'ensemble $\mathcal{F}$, un espace contenant toutes les fonctions de décision possibles qui peuvent être apprises par un modèle donné, ne contient pas la (meilleure) fonction $f_{opt} \in \mathcal{F}$ (sélectionnée par un algorithme d'apprentissage ayant la capacité de généralisation maximale) mais pouvant étendre F en combinant des classifieurs dans $\mathcal{F}_{app}$; la limitation de représentation équivaut à un *biais*.

- *Raisons computationnelles* : Outre le compromis *variance-biais*, certains algorithmes de classification sont confrontés à des problèmes de calcul. Cela est dû au fait que beaucoup d'entre eux restent bloqués dans des optima locaux lors de la recherche de la fonction de décision optimale f_{opt} dans $\mathcal{F}$. En effet, les problèmes de classification sont généralement considérés comme *NP-difficile* [43]. Cependant, un ensemble de classifieurs peut contourner / réduire le problème des optima locaux en faisant varier le point d'initialisation parmi le pool de classifieurs.

La figure 2.3 démontre les trois types de limitations, où les points noirs (•) représentent les fonctions de décision apprises, le point bleu (•) correspond à la combinaison d'éléments dans $\mathcal{F}_{app}$ et le point rouge (•) représente f_{opt}.

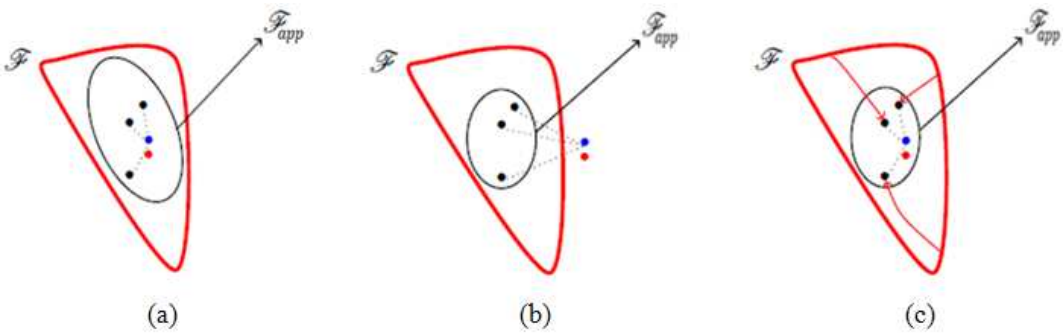

Figure 2.3 : Illustration des 3-limites pour lesquelles un ensemble est préférable à un seul classifieur : statistique (a), représentationnelles (b) et computationnelles (c) [43]

En plus de cela, il existe une autre motivation importante pour l'utilisation de la classification ensembliste, comme mentionné dans la section précédente. En effet, dans le domaine de la fusion multimodale, plusieurs ensembles de caractéristiques sont obtenus à partir de diverses sources. Ils ont cependant une nature intrinsèquement différente de sorte qu'ils ne peuvent pas être utilisés collectivement pour former un classifieur unique de manière efficace. Dans de tels cas, les caractéristiques obtenues à partir de chaque source sont utilisées pour former différents classifieurs, dont les sorties sont ensuite combinées pour prendre une décision plus éclairée/efficace.

- ✓ ***Pourquoi combiner des classifieurs?***

En outre, dans les observations suivantes et bien d'autres aussi, l'idée de faire coopérer les classifieurs est apparue :

- Il n'existe pas de *meilleur* classifieur capable de traiter (apprendre) toute distribution de données d'apprentissage.
- Aucun classifieur ne peut discriminer assez correctement un ensemble important de classes.
- Le *réglage* d'un classifieur est un problème extrêmement difficile, il se fait souvent par essais et erreurs.
- Réduire l'importance des choix initiaux.
- Tenir compte des performances de chacun des classifieurs.
- Exploiter la *complémentarité* qui peut exister entre les classifieurs.
- *La précision*, une décision plus fiable peut être obtenue en combinant les résultats de plusieurs classifieurs.
- *L'efficacité*, un problème complexe peut être décomposé en plusieurs sous-problèmes qui sont plus faciles à comprendre et à résoudre (diviser pour mieux régner), etc.

3.5.2 Que signifie un classifieur

Dans le cadre du problème de classification, un classifieur peut être défini comme un algorithme (d'apprentissage automatique) qui produit des étiquettes de classe en tant que sorties, à partir d'un ensemble de caractéristiques d'un objet. Un classifieur peut être construit directement à partir d'un ensemble d'exemples de modèles avec leurs classes respectives, ou indirectement à partir d'un modèle statistique.

On définit un ensemble de classes $\mathsf{C}_i, i = 1, \dots, N$ comprenant une distribution des objets à reconnaître. En général, un vecteur de degré d'appartenance $D_e(x)$ est associé à l'objet à reconnaître x tel que:

$$D_e(x) = \{D_e^1(x), D_e^2(x), \dots, D_e^N(x)\} \tag{2.1}$$

avec $D_e^i(x) = D_e\{x \in C_i\}$, où x peut appartenir à plusieurs classes si $D_e^i(x) \neq 0$. Néanmoins, comme la plupart des études de classification sont des problèmes de classification binaire (tels que le diagnostic des pathologies), un objet ne peut appartenir uniquement/qu'à une seule classe. D'où $D_e^i(x) = \theta_{i,j}$:

$$\theta_{i,j} = \begin{cases} 1 \; si \; i = j \\ 0 \; si \; i \; \# \; j \end{cases} \tag{2.2}$$

C_i est alors fréquemment appelée « *la vraie classe* »:

$$D_e(x) = \{0{,}0, \dots 1{,}0, \dots 0{,}0\} \tag{2.3}$$

La conception d'un classifieur implique alors la réalisation d'un *estimateur* $e(x)$ de $D_e(x)$. En règle générale, un classifieur est un système de reconnaissance (traitement des données) recevant un objet x en entrée et génère un ensemble d'informations sur la classe correspondant à cet objet. Autrement dit, en mathématique, un classifieur est considéré comme une fonction qui désigne d'affecter à x la classe C_i parmi un nombre fini de classes possible (i varie de 1 à N) en utilisant des descripteurs/caractéristiques de l'objet x à reconnaître pour lesquelles il n'y a pas d'information a priori [44, 45]:

$$x \rightarrow^{e(x)} = \mathsf{C}_i \tag{2.4}$$

Le classifieur peut être employé dans n'importe quel domaine d'application. D'une manière générale, le choix d'une représentation pour décrire les données/caractéristiques, d'un algorithme de décision et d'un ensemble d'apprentissage permettant de régler les paramètres du classifieur qui est considéré comme une phase indispensable (dans un premier temps) pour la mise en place de tout classifieur. Les performances de ce dernier peuvent être modifiées en introduisant des ajustements soit au niveau des caractéristiques qu'il traite, soit au niveau de ses paramètres (type de sorties, règles de décision, etc.).

Dans le contexte de la combinaison, un classifieur e_k peut être signifié comme étant un outil de reconnaissance d'objets qui opère dans une partie spécifique de caractéristiques, qui adopte un certain ensemble de données pour l'apprentissage, qui prend sa décision sur la base d'une certaine règle et qui génère un certain type de réponse en sortie.

En outre, on note $E = \{(x_i, y_i), i \in [1, \dots, N]\}$ comme l'ensemble d'apprentissage dans lequel y_i est la véritable étiquette de x_i, où $y_i \in Y = \{\mathsf{C}_m\}, m = 1, \dots, M$; et k comme le nombre de classifieurs de base. Pour une observation x_i, le k^{th} classifieur renvoie les réponses, c'est-à-dire les prédictions selon lesquelles cette observation appartient aux étiquettes de classe du jeu d'étiquettes Y. En fait, ces réponses peuvent être considérées comme une estimation des probabilités postérieures que x_i appartient aux étiquettes de classe. $R_{k,m}(x_i)$ est désigné comme réponse du k^{th} classifieur pour la m^{th} classe sur x_i dans laquelle $R(x_i) \in [0,1]$ et $\sum_m R_{k,m}(x_i) = 1$ [46].

Sur la base du niveau d'informations fournies par le classifieur et selon la catégorisation utilisée dans la plupart des études concernant la combinaison des classifieurs [25, 47, 48], la réponse produite par le classifieur e_k pour un objet d'entrée x peut être répartie en trois types principaux (P1):

a. *Type classe :* Dans cette catégorie, seule l'étiquette de classe C_i assignée par le classifieur e_k à l'objet x est fournie :

$$e_k(x) = \mathsf{C}_i, o\grave{u}\ i \in \{1, \dots, N\} \tag{2.5}$$

Dans ce cas, l'opinion du classifieur est représentée sous forme binaire ; il peut également générer un pool de classes. Le classifieur considère ensuite que l'objet x appartient à l'une des classes de ce pool sans procurer aucune autre information supplémentaire permettant de distinguer les classes.

b. *Type rang :* Dans ce type, le classifieur e_k assigne le rang $r_{i,k}$ à la classe C_i pour l'objet x :

$$e_k(x) = (r_{1,k}, r_{2,k}, \dots, r_{N,k}) \tag{2.6}$$

Il s'agit d'un vecteur des rangs de taille N triés par ordre décroissant de préférence/plausibilité à propos des classes, produit par le classifieur (comme résultat) indiquant un classement sur les classes. Ainsi, la classe ayant le rang le plus élevé dans

la liste suggérée par le classifieur représente la classe la plus probable pour l'objet x et la classe avec le rang le plus bas est la moins probable.

c. *Type mesure :* Ce type signifie que la mesure $m_{i,k}$ est affectée à la classe C_i pour l'objet x par l'intermédiaire du classifieur e_k :

$$e_k(x) = (m_{1,k}, m_{2,k}, \dots, m_{N,k}) \tag{2.7}$$

Ce type d'information reflète le niveau de confiance du classifieur dans sa suggestion. Par conséquent, un vecteur de mesures (ou *de certitudes*) de taille N est fourni en tant que sortie du classifieur. Cette mesure de confiance, standardisée ou non, peut être définie sous plusieurs formes, à savoir: une mesure de similitude, une mesure de distance, une probabilité a posteriori, une valeur de confiance, un score, une fonction de croyance, etc.

Semblable de ce qui a été discuté dans les sous-sections précédentes, le classifieur génère un niveau différent d'informations correspondant à chaque type de sortie (*classe, rang ou mesure*). La sortie du type de classe représente le type le plus simple à mettre en œuvre mais avec le moins d'informations. La sortie du type rang signifie l'ordre de préférence des suggestions procurées par le classifieur ayant des informations moins importantes en les comparant au type mesure. La sortie du type mesure fournit les informations les plus riches par rapport aux autres types de sortie (classe et rang) car elle indique le niveau de confiance du classifieur dans ses propositions.

✓ ***Classifieur de base***

La classification d'ensemble est composée d'un ensemble de classifieurs de base représentant les composants de base d'un *système multi-classifieurs*. Les classifieurs de base doivent être différents, mais ils doivent également être comparables afin que leurs sorties puissent être combinées.

Dans de nombreuses études, une transformation des sorties des classifieurs de base du type rang ou mesure au type classe (avec perte d'informations) est requise en vue d'uniformiser leurs sorties pour des scénarios particuliers. Cela consiste à ne prendre en compte que la meilleure solution de la liste suggérée par chaque classifieur. En d'autres termes, pour le type rang, il suffit de sélectionner la classe qui occupe la première place représentant le rang le

plus élevé ; concernant le type mesure, il suffit de sélectionner la classe ayant la meilleure valeur/mesure.

3.5.3 Évaluation des performances de classification

La classification est une approche d'apprentissage supervisé dans laquelle une variable cible est catégorique (ou discrète). L'évaluation d'un classifieur/modèle d'apprentissage automatique est aussi importante que sa conception. Les modèles sont créés pour fonctionner sur de nouvelles données non inédites. Par conséquent, une évaluation approfondie et polyvalente est nécessaire pour créer un modèle ou former un classifieur robuste. Une gamme de différentes métriques d'évaluation (comme *la matrice de confusion*) avec leurs avantages et leurs inconvénients a été suggérée dans la littérature pour différents types de problèmes de classification en vue de représenter la performance d'un classifieur et de vérifier son applicabilité [44, 49].

La principale façon d'évaluer les performances des classifieurs consiste à dériver leurs probabilités fréquentistes de succès sur un ensemble de validation Ω_{val}. Un ensemble de validation contient également des paires $\{(x_i, y_i)\}$ qui ne sont pas exploitées pendant la phase d'apprentissage ($\Omega_{val} \cap \Omega_{app} = \emptyset$). La taille de l'ensemble de validation est indiquée par Ω_{val} (le choix de la taille de l'ensemble Ω_{val} se fait au détriment de l'ensemble d'apprentissage Ω_{app}). Il existe diverses méthodes de validation d'un modèle de classification qui ont été décrites par Boser et al. [50]; ce manuscrit en révèle certaines.

- La validation *holdout* est la méthode la plus courante et la plus simple pour évaluer un classifeur. Dans cette méthode, l'ensemble de données spécifié est divisé (au hasard) en deux partitions: ensemble d'apprentissage et ensemble de test/validation. En règle générale, l'ensemble de données d'apprentissage est plus grand que l'ensemble de données de validation. Les ratios typiques utilisés pour diviser l'ensemble de données comprennent 70:30, 80:20, etc. La limitation d'une telle méthode est que l'erreur trouvée dans l'ensemble de données de test peut dépendre fortement des observations incluses dans l'ensemble de données d'Ω_{app} et d'Ω_{val}. De plus, si l'Ω_{app} ou l'Ω_{val} ne sont pas en mesure de représenter les données complètes réelles, les résultats des ensembles de test peuvent être faussés/biaisés.
- La *validation croisée k-blocs* « *k-fold cross-validation* » est un moyen d'améliorer la méthode et peut être effectuée pour vérifier que le modèle n'est pas en état de *sur-apprentissage*. Dans cette méthode, l'ensemble de données est partitionné de manière

aléatoire en k sous-ensembles mutuellement exclusifs, chacun de taille approximativement égale et un est conservé pour les tests tandis que d'autres sont utilisés pour l'apprentissage. Ce processus est répété dans l'ensemble des k blocs (autrement dit, la méthode *holdout* est répétée k fois). La valeur de k est généralement égale à 5 ou 10.

- Lorsque la valeur de k est fixée à N, où N représente la taille de l'ensemble de données, l'approche est connue sous le nom de « *leave-one-out cross-validation* » (cas particulier de la validation croisée k-blocs), afin de donner à chaque échantillon de test une opportunité d'être utilisé dans l'ensemble de données *holdout*.

Les probabilités dérivées pour un classifieur de l'ensemble de validation sont des estimations des probabilités conditionnelles $p(\hat{c} = c_j|Y = c_i)$. Toutes ces probabilités sont dérivées de la matrice de confusion M du classifieur:

$$M = \begin{bmatrix} \sum_{i=1}^{n_{val}} \pi_{C_1}(y^{(i)})\pi_{C_1}(\hat{c}(x^{(i)}) & \cdots & \sum_{i=1}^{n_{val}} \pi_{C_1}(y^{(i)})\pi_{C_m}(\hat{c}(x^{(i)}) \\ \vdots & & \vdots \\ \sum_{i=1}^{n_{val}} \pi_{C_m}(y^{(i)})\pi_{C_1}(\hat{c}(x^{(i)}) & \cdots & \sum_{i=1}^{n_{val}} \pi_{C_m}(y^{(i)})\pi_{C_m}(\hat{c}(x^{(i)}) \end{bmatrix},$$

Où π_{C_i} est la fonction indicatrice de la classe C_i. Les lignes et les colonnes font respectivement référence aux classes réelles et prévues. L'élément M_{ij} est le nombre d'exemples dont la véritable étiquette (la classe réelle) est C_i mais qui ont été classés comme C_j par le classifieur. Le nombre $n_{val,i}$ d'exemples de classe C_i dans l'ensemble de validation est donné par la somme des éléments de la i^{th} ligne. La normalisation de chaque ligne par ces nombres donne les probabilités estimées:

$$\hat{p}(\hat{c} = c_j|Y = c_i) = \frac{M_{ij}}{\sum_{j=1}^{m} M_{ij}} = \frac{M_{ij}}{n_{val,i}} \tag{2.8}$$

Les critères de performance de la classification standard sont également calculables à partir de $M^{(k)}$. La précision est définie comme suit :

$$précision_i = \frac{M_{ii}}{\sum_{j=1}^{m} M_{ij}} = \frac{M_{ii}}{n_{val,i}} \tag{2.9}$$

Le taux de précision de la classe C_i est interprété comme la probabilité qu'un exemple choisi au hasard dont la classe prédite est C_i possède une véritable étiquette C_i.

De même, le rappel « *recall* » de la classe C_j est la probabilité qu'un exemple sélectionné aléatoirement dont la véritable étiquette est C_j, soit prédit comme C_j:

$$rappel_j = \frac{M_{jj}}{\sum_{i=1}^{m} M_{ij}} = \frac{M_{jj}}{n_{val,j}} \tag{2.10}$$

Enfin, la précision globale « *accuracy* » du classifieur est fournie par :

$$accuracy = \frac{\sum_{i=1}^{m} M_{ii}}{\sum_{i=1}^{m}\sum_{j=1}^{m} M_{ij}} = \frac{\sum_{i=1}^{m} M_{ii}}{n_{val}} \tag{2.11}$$

Une autre mesure intéressante pour l'évaluation du classifieur est la fonction d'efficacité du récepteur, plus communément désignée sous le nom « *courbe ROC* ». Les courbes ROC (*Receiver Operating Characteristic* » sont également un outil d'évaluation important d'un modèle de classification [51]. Une courbe ROC est un tracé graphique illustrant les performances d'un classifieur binaire et peut être utilisé pour comparer un ensemble de classifieurs. Avant d'expliquer la courbe ROC, les concepts de *sensibilité* et de *spécificité* dans les tâches de classification d'apprentissage automatique doivent être clarifiés. Compte tenu d'un problème de classification binaire $\Omega = \{C_1, C_2)$, la *sensibilité* et la *spécificité* sont respectivement les valeurs de rappel pour les classes C_1 et C_2:

$$sensibilité = \frac{M_{11}}{M_{11} + M_{21}} \tag{2.12}$$

$$spécificité = \frac{M_{22}}{M_{22} + M_{12}} \tag{2.14}$$

La courbe ROC montre le tracé de $(1 - spécificité)$ sur l'axe horizontal par rapport à la sensibilité sur l'axe vertical et donc un ensemble de points $(1 - spécificité,\ sensibilité)$ est requis. Lors de la construction de la matrice de confusion, une probabilité de seuil de coupure régulière de 0,5 est sélectionnée (Ce seuil est celui correspondant à la règle de classification reposant sur $arg\ max_{y\in\Omega}\ p(Y = y|x)$ pour les classifieurs probabilistes), ce qui signifie qu'un échantillon est attribué à C_i si $p(Y = C_i|x) > 0.5$ et donc une paire $(1 - spécificité, sensibilité)$ est obtenue. Lorsque le coût d'une mauvaise classification des exemples de la classe C_1 n'est pas le même que le coût d'une mauvaise classification d'un exemple de la classe C_2, d'autres valeurs de seuil sont utilisées et un ensemble de points permettant de construire la courbe ROC est obtenu. Une autre mesure qui peut être extraite du graphique ROC est $l'aire\ sous\ la\ courbe\ ROC$ « AUC ». *L'AUC* est alors la probabilité qu'un exemple choisi au hasard dans la classe C_2 obtienne une valeur de s (un score retourné par le classifieur pour la classe C_2) supérieure à celle obtenue par un exemple choisi au hasard dans la classe C_1.

3.5.4 Topologies de combinaison

En vue de construire un système multi-classifieurs et de combiner les différentes sorties de ces derniers pour définir une décision finale unique, divers scénarios et stratégies de combinaison sont introduits dans la littérature, à savoir *séquentielle*, *parallèle* et *hybride* (P2).

- ***Topologie séquentielle***

La combinaison séquentielle, également connue sous le terme de combinaison en *cascade* ou en *série*, est structurée en étapes décisionnelles consécutives dans le but de diminuer progressivement le nombre de classes potentielles. En d'autres termes, chaque phase implique un classifieur unique qui prend en considération la réponse donnée par le classifieur situé en amont (*le résultat de la classification généré par un classifieur est utilisé comme entrée dans le classifieur suivant*) afin de traiter les rejets ou confirmer la décision obtenue sur l'objet x (à identifier) qui lui est présenté.

Grâce à ce processus, un problème complexe est progressivement réduit à un problème plus simple (figure 2.4).

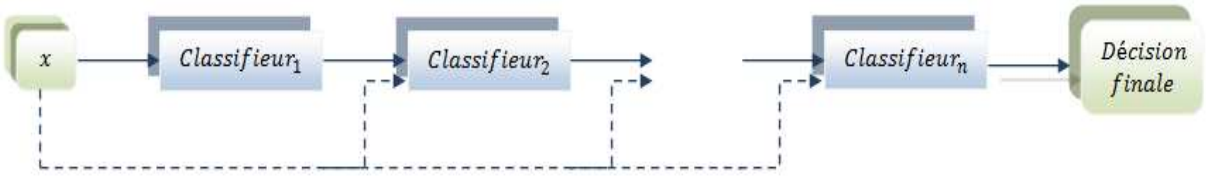

Figure 2.4 : Combinaison séquentielle de classifieurs

Toutefois, l'ordre dans lequel les classifieurs sont placés est très important pour ce type de combinaison, le premier classifieur doit être choisi judicieusement afin d'éviter - autant que possible - de propager la mauvaise réponse.

Un exemple de cette architecture est la généralisation en cascade présentée par Gama and Brazdil [52]. L'idée de base derrière cette approche est d'utiliser des classifieurs en séquence, à la différence d'utiliser une extension des données d'apprentissage pour les niveaux supérieurs. À chaque itération, de nouveaux attributs sont ajoutés aux modèles d'apprentissage. Ces nouveaux attributs représentent la probabilité que les modèles appartiennent à des classes spécifiques. La couche unique au niveau final prendra la décision finale.

- ***Topologie parallèle***

Alternativement, dans une architecture parallèle, l'ordre dans lequel les classifieurs sont exécutés n'est pas important. Les classifieurs (de base) fonctionnent d'abord indépendamment les uns des autres (*en parallèle*) pour prédire l'étiquette d'un objet inconnu x. Par la suite, leurs décisions individuelles sont combinées pour aboutir à une décision unique (figure 2.5).

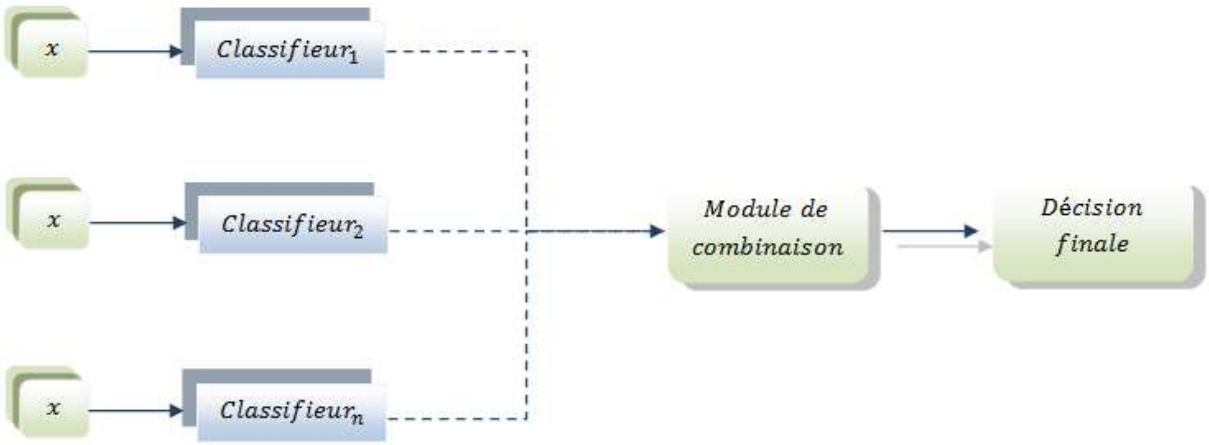

Figure 2.5 : Combinaison parallèle de classifieurs

Dans ce type de stratégie, différents classifieurs sont formés par le même ensemble de caractéristiques ou bien le même classifieur est formé par différentes familles de caractéristiques. C'est la topologie la plus fréquemment utilisée dans les systèmes multi-classifieurs et elle a été étudiée, tant empiriquement que théoriquement en raison de ses résultats impressionnants et de sa simplicité de mise en œuvre [36, 37, 25, 53]. Safont et al. [54] suggèrent de combiner quatre classifieurs simples (fonctionnant en parallèle), à savoir, l'analyse discriminante linéaire (ADL) et quadratique (ADQ), Bayes naïfs (BN) et forêt aléatoire (FA) et trois méthodes de fusion compétitives ont été estimés pour comparaison, telles que la moyenne, le vote majoritaire et l'intégration séparée des scores (ISS). L'idée de cette étude est d'avoir réussi à combiner de manière optimale les résultats de tous les classifieurs individuels en utilisant l'intégration des scores vectoriels (VSI).

- ***Topologie hybride***

La combinaison hybride consiste en un mélange d'architectures séquentielles et parallèles en vue de tirer pleinement et simultanément profit de deux approches. La figure 2.6 illustre un exemple de cette combinaison dans laquelle un classifieur est agrégé en série avec deux classifieurs en parallèle, d'une part en diminuant l'ensemble des classes possibles et d'autre

part en ajoutant un module d'agrégation pour combiner les différentes sorties de classifieurs afin de fournir la décision finale.

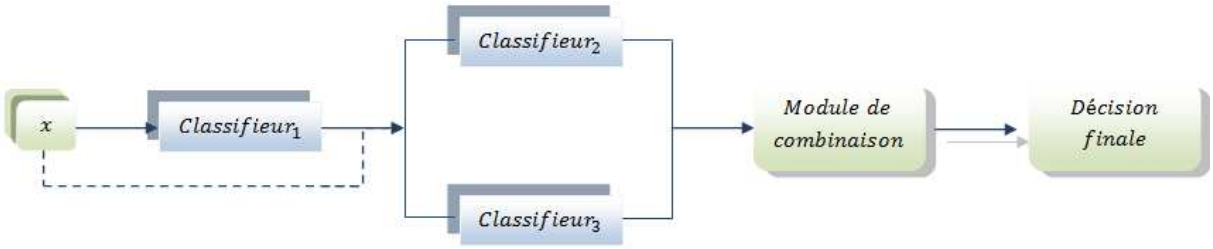

Figure 2.6 : Combinaison hybride de classifieurs

Dans ce mode de fonctionnement, les classifieurs individuels sont combinés dans une structure arborescente. Ce mode est préférable lorsqu'un grand nombre de classifieurs sont attendus.

D'autres chercheurs ont suggéré d'adopter une structure *conditionnelle*, où un classifieur primaire est d'abord appliqué aux données d'entrée. Si la classification est effectuée avec un faible niveau de confiance, un autre classifieur est utilisé. Cette architecture présente l'avantage d'une efficacité de calcul lorsque le classifieur initial est un type à faible coût et que le second peut être plus sophistiqué pour les modèles difficiles. Asker et Maclin [55] propose une méthode qui repose sur une simple estimation des performances de chaque classifieur. Les classifieurs sont regroupés dans une liste ordonnée où chaque classifieur a un seuil correspondant. Lors de la phase de classification, le premier classifieur de la liste est consulté. Si la confiance de prédiction de ce classifieur est supérieure au seuil prédéfini, ce classifieur est alors utilisé pour la prise de la décision finale. Sinon, le classifieur suivant et son seuil sont pris en compte. Si aucune des prédictions faites par les classifieurs n'est supérieure au seuil, une technique de calcul de la moyenne sera effectuée sur toutes les prédictions.

Toutefois, les topologies séquentielles et hybrides ont été largement négligées par les scientifiques. La plupart des recherches dans le domaine de la combinaison de classifieurs se focalisent essentiellement sur la combinaison parallèle en utilisant un ensemble identique/différent de données d'entrée et sur la fusion de leurs résultats pour aboutir à une décision unique. De nombreuses motivations justifient l'intérêt manifesté par la majorité des chercheurs pour la combinaison parallèle de classifieurs, à savoir :

- Dans la combinaison, il est envisageable d'employer diverses famille de caractéristiques de dimension importantes mais en les répartissant sur différents classifieurs.
- La combinaison de l'ensemble/parallèle peut prendre en compte les avantages des classifieurs appris d'une manière différente au lieu de rechercher la meilleure configuration de ses paramètres en raison de la sensibilité (généralement) aux choix initiaux des paramètres de ces modèles.
- L'apprentissage d'un classifieur unique sur plusieurs bases d'apprentissage, chacune étant collectée différemment ou construite dans des conditions différentes, peut générer des résultats divers.
- Dans le pool, les classifieurs ont leurs propres régions (séparation) dans l'espace des caractéristiques où ils sont les plus efficaces, mais peuvent avoir des performances globales similaires.

3.5.5 Catégorisations des méthodes de combinaison parallèle

Dans un système multi-classifieurs basé sur le principe de la combinaison parallèle, où chaque classifieur fonctionne indépendamment sur le même objet à reconnaître/classer. Le principal défi est de savoir comment élaborer une décision finale unique à partir des sorties fournies par ces modèles? Ce défi consiste à utiliser un opérateur ou une règle de combinaison afin d'obtenir un résultat final.

La variété des techniques d'ensemble a émergé plusieurs taxonomies dans la littérature [22, 25, 48, 54], qui visent à catégoriser les méthodes de combinaison sous différents points de vue. Xu et al. [25] ont montré que les méthodes de combinaison ne se distinguent que par le type de sorties des classifieurs (*classe*, *rang*, *mesure*) fournis en entrée pour la combinaison. Le type de sorties des classifieurs (de base) est un facteur qui figure dans quasiment toutes les taxonomies mentionnées dans la littérature. Les méthodes de la classification d'ensemble utilisent un type de niveau de sortie différent. En général, deux stratégies de combinaison de classifieurs sont identifiées: la fusion et la sélection de classifieurs [56]. Rappelons que la fusion réside dans la combinaison de toutes les sorties des classifieurs pour obtenir un résultat unique, tandis que la sélection revient à choisir statiquement/dynamiquement parmi un pool de classifieurs potentiels, les meilleurs classifieurs (les plus complémentaires) pour reconnaître l'objet inconnu.

En outre, la catégorisation des méthodes de combinaison parallèle (fusion) se caractérise par la nature des classifieurs à combiner, à savoir: des classifieurs *hétérogènes*, c'est-à-dire que les apprenants de base sont de types distincts (en d'autres termes, les méthodes d'ensemble hétérogènes visent à combiner un ensemble d'hypothèses $h_1 \dots h_T$ produites par différents algorithmes $L_1 \dots L_T$ sur le même ensemble d'apprentissage) et des classifieurs *homogènes*, également appelés des apprenants de base faibles qui ont la même architecture (principale) mais sont formés sur des données/caractéristiques différentes ou configurés différemment [57] (autrement dit, les méthodes d'ensemble homogènes ont pour but de combiner un ensemble d'hypothèses $h_1 \dots h_T$ générées au moyen du même algorithme $L_1 \dots L_T$ appliqué à un ensemble d'apprentissage différent) (P3).

En effet, la combinaison parallèle peut être mise en œuvre en utilisant différentes stratégies, la capacité d'apprentissage et le type d'informations générées par les classifieurs [58]. Comme déjà indiqué, les schémas de combinaison peuvent être divisés en deux catégories, notamment les méthodes $sans - apprentissage$ et les méthodes $avec - apprentissage$, qui peuvent varier de simples, comme les règles d'agrégation linéaire (telles que la somme, la médiane, le maximum, le minimum, le produit et le vote à la majorité) à des techniques plus complexes, comme les règles pondérées (somme et produit) et le vote pondéré (P4).

3.5.5.1 Combinaison sans-apprentissage

Cette catégorie est également appelée *non-paramétrique* (ne nécessitant pas d'étape d'apprentissage) et est communément adoptée en raison de la simplicité de sa mise en œuvre. Néanmoins, les méthodes de cette combinaison n'utilisent que les sorties des classifieurs qui les traitent de manière identique, ce qui ne permet pas de prendre en compte leur capacité particulière. Cette combinaison est répartie en type classe, rang et mesure. Dans la suite de cette section, quelques règles de méthodes de combinaison bien connues sont présentées.

a) Type Classe

Lors de la combinaison des sorties d'étiquettes de classifieur, le moyen le plus simple (évident) est de recourir à un système de vote. Il présente l'avantage de pouvoir être employée pour n'importe quel type de classifieur. Plusieurs variantes de schémas de vote peuvent pratiquement toutes être déduites de la $règle\ avec\ seuil$ définie par :

$$E(x) = \begin{cases} C_i & si \ \sum e_{i,j} = max_{t=1}^{n} \sum_{j=1}^{L} e_{t,j} \geq \gamma \cdot L \\ rejet & sinon \end{cases} \quad (2.15)$$

où γ correspond au $seuil$ représentant le ratio du nombre de classifieurs devant répondre à la même classe pour que cette classe soit la classe finale de la combinaison sinon rejet.
En général, il existe trois techniques de vote en fonction de la valeur du seuil γ.

- *Vote à la pluralité*

Pour la valeur du seuil $\gamma = 0$, c'est-à-dire que la classe ayant le plus grand nombre de votes est désignée comme classe/décision finale. Il est cependant rejeté si toutes les classes ont le même nombre de votes. Dans ce cas, les risques de confrontation sont particulièrement élevés.

- *Vote à la majorité absolue*

Pour la valeur du seuil $\gamma = 0.5$, la détermination de la classe finale nécessite de réunir plus de la moitié des décisions des classifieurs concernant la classe en question. Autrement dit, si au moins Q classifieurs sont du même avis, Q peut être décrit comme suit :

$$Q = \begin{cases} \frac{L}{2} & si\ L\ est\ pair \\ \frac{L+1}{2} & si\ L\ est\ impair \end{cases} \quad (2.16)$$

- *Vote à l'unanimité*

Pour la valeur du seuil $\gamma = 1$, la sélection de la classe finale nécessite que tous les classifieurs suggèrent cette décision sinon rejet.

b) Type rang

Les approches de type rang combinent des listes de suggestions de classifieurs triées dans un ordre décroissant selon leur préférence. Les méthodes non-paramétriques de ce type les plus répandues sont le *Borda count* et le *meilleur rang*.

- *Borda Count*

Borda Count consiste en un schéma de combinaison basé sur le classement dans lequel chaque classifieur hiérarchise les classes (candidats) en fonction de leurs chances d'être la bonne/vraie classe [59]. Chaque rang associé à un score allant de $n-1$ pour le premier rang à 0 pour le dernier où n est le nombre total de classes. Dans un deuxième temps, la somme des scores reçus par chaque classe est calculée, et l'étiquette de classe obtenant le score cumulé le plus élevé représente la prédiction d'ensemble:

$$BC(\mathsf{C}_i) = \sum_{j=1}^{L} r_{i,j} \tag{2.17}$$

où $r_{i,j}$ est le rang affecté par le classifieur e_j à la classe C_i.

- *Meilleur rang*

L'idée derrière cette méthode est d'assigner à chaque classe le rang le plus élevé $MR\ (C_i)$ parmi les rangs proposés par les classifieurs et d'ordonner la liste en fonction de ces rangs.

$$MR\ (C_i) = \max_{j\,=\,1}^{L} r_{i,j} \tag{2.18}$$

c) Type mesure

Les techniques de cette catégorie combinent des scores/mesures correspondant au niveau de certitude des classifieurs en termes d'appartenance de l'objet à identifier dans chacune des classes. Néanmoins, étant donné que les sorties des classifieurs ne sont pas toujours semblables, une standardisation est souvent inévitable. Les règles fixes représentant les méthodes de combinaison *non-paramétriques* les plus courantes sont décrites ci-après. En effet, afin de pouvoir identifier un objet x, une règle de décision $E(x)$ est utilisée qui équivaut à sélectionner la classe C_i pour laquelle la probabilité a posteriori P_i est la plus élevée:

$$E(x) = \begin{cases} \mathsf{C}_i & si \quad max_{i=1}^{n} P_i = max_{m=1}^{n} P_m \\ rejet & sinon \end{cases} \tag{2.19}$$

Cette probabilité a posteriori P_m est déterminée selon l'une des règles ci-dessous [57]:

- *La règle linéaire*

$$P_m = \gamma \sum_{j=1}^{L} m_{i,j} \tag{2.20}$$

où, γ est une constante. Ainsi, *la règle de somme* est obtenue si $\gamma = 1$, pour $\gamma = \frac{1}{L}$, la règle correspond alors à *la moyenne simple*.

- *La règle produit*

$$P_m = \prod_{j=1}^{L} m_{i,j} \tag{2.21}$$

- *La règle médiane*

$$P_m = \begin{cases} \dfrac{m_{i,\frac{L}{2}} + m_{i,\frac{L+2}{2}}}{2} & si\ L\ est\ pair \\ m_{i,\frac{L+1}{2}} & si\ L\ est\ impair \end{cases} \tag{2.22}$$

- *La règle maximum*

$$P_m = max_{j=1}^{L} m_{i,j} \tag{2.23}$$

- *La règle minimum*

$$P_m = min_{j=1}^{L} m_{i,j} \tag{2.24}$$

3.5.5.2 Combinaison avec-apprentissage

Par opposition aux approches précédemment citées, les méthodes de combinaison $avec-apprentissage$ ou $paramétrique$ sont plus délicates à mettre en œuvre. Ces techniques recourent à des paramètres complémentaires élaborés lors d'une étape d'apprentissage. Dans la combinaison $avec-apprentissage$, il est souvent souhaitable de se servir d'un deuxième ensemble de données au stade de la combinaison.

a) Type classe

Concernant ce type de combinaison, le vote pondéré, la théorie de Bayes et la méthode d'espace de connaissance du comportement (Behaviour Knowledge Space ou BKS) constituent essentiellement les méthodes de combinaison *paramétrique* les plus répandues.

- *Vote avec pondération*

Dans ce schéma de vote, la sortie $e_{i,j}$ de chaque classifieur e_j est pondérée par un coefficient w_j désignant son poids dans la combinaison.

$$E(x) = \begin{cases} \mathsf{C}_i & si \ \sum_{j=1}^{L} w_j e_{i,j} = max_{k=1}^{n} \sum_{j=1}^{L} w_j e_{k,j} \\ rejet & sinon \end{cases} \quad (2.25)$$

où, les coefficient w_j peuvent être déterminés soit en les optimisant via *l'algorithme génétique* [60], soit simplement en utilisant des votes pondérés par la fiabilité ($\frac{taux\ de\ reconnaissance}{100-taux\ de\ rejet}$) estimée de chacun des classifieurs.

- *Théorie de Bayes*

Le recours à la règle bayésienne [61] revient à définir la classe C_i pour laquelle la probabilité a postérieure $P(\mathsf{C}_i/e_1 = \mathsf{C}_1, \dots, e_L = \mathsf{C}_L)$ est maximale, à savoir:

$$E(x) = \begin{cases} \mathsf{C}_i \ si \ \ P(\mathsf{C}_i/e_1 = \mathsf{C}_1, \dots, e_L = \mathsf{C}_L) \ = max_{m=1}^{n} P(\mathsf{C}_m/e_1 = \mathsf{C}_1, \dots, e_L = \mathsf{C}_L) \\ rejet \ \ sinon \end{cases} \quad (2.26)$$

Selon l'hypothèse de l'indépendance, la probabilité a postérieure s'écrit:

$$P\left(\mathsf{C}_i/e_1 = \mathsf{C}_1, \dots, e_L = \mathsf{C}_L\right) = P(\mathsf{C}_i) \prod_{l=1}^{L} \frac{P(\mathsf{C}_i/e_l = \mathsf{C}_l)}{P(\mathsf{C}_i)} \quad (2.27)$$

Les probabilités peuvent être calculées à partir des matrices de confusion :

$$P(\mathsf{C}_i/e_l = \mathsf{C}_l) = \frac{n^j_{\mathsf{C}_i,\mathsf{C}_l}}{n^j_{.,\mathsf{C}_l}} \quad (2.28)$$

où, $n^j_{\mathsf{C}_i,\mathsf{C}_l}$ représente le nombre de données pour lesquelles le classifieur e_j affecte des données de la classe C_i à la classeC_l et $n^j_{.,\mathsf{C}_l}$ est le nombre total de données assignées à la classe C_l.

- *Méthode d'espace de connaissance du comportement (BKS)*

Chaque combinaison possible d'étiquettes de classe est un index qui est représenté par une cellule dans le tableau de correspondances. L'étiquette de classe le plus souvent rencontré parmi les éléments de ce tableau lors de l'apprentissage est sélectionnée pour la cellule. Les décisions générées par chaque classifieur sont comparées à ce tableau de correspondances. La décision d'un objet inconnu est prise en fonction de la classe stockée dans la cellule [62].

b) Type rang

Les méthodes *non-paramétriques* de ce type de combinaison indiquées ci-dessus traitent les performances des classifieurs de la même manière. En d'autres termes, tous les classifieurs interviennent de manière égale dans la prise de décision, même s'il existe une supériorité entre eux. Il est donc primordial de tenir compte du degré de fiabilité des classifieurs dans la combinaison. La somme pondérée des rangs et la méthode de régression logistique représentant les règles de type rang largement utilisées sont brièvement décrites comme suit:

- *Somme pondérée*

Dans cette règle, les rangs assignés par les classifieurs pour une classe C_i sont pondérés par des coefficients w_j désignant la fiabilité de chaque classifieur e_j. La somme pondérée des rangs (SPR) d'une classe C_i est définie comme suit:

$$SPR(\mathsf{C}_i) = \sum_{j=1}^{L} w_j r_{i,j} \tag{2.30}$$

- *Régression logistique*

Une autre façon de combiner les classifieurs de type rang est de recourir à la méthode de régression logistique qui consiste à utiliser les poids en tant qu'informations indiquant l'importance relative des classifieurs. Cette technique est fondée sur $\pi(\mathsf{C}_i)$, la probabilité de la classe C_i avec $0 \leq \pi(\mathsf{C}_i) \leq 1$. La fonction *logistique* est adoptée afin de calculer cette probabilité qui s'écrit :

$$\pi(\mathsf{C}_i) = \frac{\exp\left(\alpha + \beta_1 r_{i,1} + \beta_2 r_{i,2} + \cdots + \beta_L r_{i,L}\right)}{1 + \exp\left(\alpha + \beta_1 r_{i,1} + \beta_2 r_{i,2} + \cdots + \beta_L r_{i,L}\right)} \tag{2.31}$$

$r_{i,j}$ représente le rang attribué par le classifieur e_j à la classe C_i. Les paramètres de régression α et β peuvent être calculés en utilisant les méthodes des *moindres carrés* ou du *maximum de vraisemblance* [63].

c) *Type mesure*

Dans le contexte des méthodes paramétriques concernant ce type, la combinaison peut faire appel à des approches de classification communes pour fusionner les classifieurs de type mesure telles que les règles pondérées (la moyenne et le produit) et l'intégrale floue [64].

- *La moyenne pondérée*

Cette méthode est semblable au vote de la moyenne simple, à l'exception du fait que la sortie des classifieurs de base est multipliée par un poids. La combinaison de divers classifieurs est construite en formant des sommes pondérées des sorties des classifieurs. La probabilité a posteriori P_i d'une classe C_i estimée en utilisant la règle suivante :

$$P_i = \frac{1}{L}\sum_{j=1}^{L} w_j m_{i,j} \tag{2.32}$$

Les poids de combinaison, w_i, sont obtenus en minimisant l'erreur quadratique moyenne des classifieurs sur les données d'apprentissage.

- *Le produit pondéré*

Cette technique applique le même principe que la règle précédente, c'est-à-dire que les sorties des classifieurs de base sont pondérées par un poids. La règle ci-après détermine la probabilité a posteriori P_i d'une classe C_i :

$$P_i = \prod_{j=1}^{L} m_{i,j}^{w_j} \tag{2.33}$$

- *L'intégrale floue*

Le concept de l'intégrale floue consiste à déterminer les mesures floues $H = \{h_{i,1}, h_{i,2}, \dots, h_{i,L}\}, i = 1, \dots, N$ pour chacune des classes ($N\ classes$), afin de les comparer aux résultats des classifieurs pour conserver la classe avec la mesure floue la plus élevée. Pour chaque classe $\mathsf{C}_i, t \in \{1, \dots, L\}$, les mesures floues $h_{i,t}$ sont estimées comme suit :

$$h_{i,t} = g^t + h_{i,t-1} + \gamma g^t h_{i,t-1} \tag{2.34}$$

avec $h_{i,1} = g^1$, γ représente une mesure floue$\gamma \geq -1$ obtenue à partir des densités floues $g^j, j = 1$ à L par la résolution de l'équation suivante :

$$\gamma + 1 = \prod_{j=1}^{L}(1 + \gamma g^{j}) \tag{2.35}$$

Pour chaque classe, les résultats des classifieurs sont comparés avec le vecteur de H en vue de définir la mesure floue f la plus élevée:

$$f_i = max_{i=1}^{N}\left[min_{t=1}^{L}(h_{i,t}, e_{i,t}\right] \tag{2.36}$$

4 Conclusion

Les informations dans le monde réel se présentent généralement sous différentes modalités. Différentes modalités se caractérisent par des propriétés statistiques très différentes. En raison des propriétés statistiques distinctes des différentes ressources d'information, il est très important de déterminer la relation entre les différentes modalités.

L'apprentissage multimodal est un modèle intéressant pour représenter les représentations communes des différentes modalités. Le modèle d'apprentissage multimodal est également capable de combler les modalités manquantes compte tenu de celles observées.

La classification à l'aide de données multimodales se présente dans de nombreuses applications d'apprentissage automatique. Il est crucial non seulement de modéliser efficacement les relations intermodales, mais également d'assurer la robustesse contre la perte d'une partie des données ou des modalités.

La tâche de classification multimodale du monde réel peut être suffisamment complexe pour un seul modèle d'apprentissage automatique. L'un des moyens possibles d'obtenir des résultats plus précis réside dans l'utilisation de plusieurs modèles (apprentissage d'ensemble) pour traiter plusieurs modalités (par exemple, la modalité d'image et la modalité textuelle) de la même source de données.

Une étude détaillée en ce qui concerne la classification multimodale en utilisant l'apprentissage d'ensemble a été menée dans ce chapitre tout en discutant de la valeur de la *multimodalité* pour les systèmes multi-classifieurs.

Ce chapitre a présenté le concept d'un système de fusion multimodal en illustrant les différents niveaux de fusion de celui-ci ainsi que les approches utilisées pour chaque niveau de fusion (capteur, caractéristique, score, rang et décision) en mettant l'accent sur les niveaux supérieurs tels que score, rang et décision *« combinaison de classifieurs »*.

Dans le cadre de la combinaison de classifieurs, les différentes topologies de combinaison ont été définies en se concentrant sur la combinaison parallèle et ses méthodes de fusion qui se distinguent principalement par le type de sortie des classifieurs. En effet, la combinaison parallèle de classifieurs est couramment utilisée en raison de sa simplicité de mise en œuvre et ses performances pour diverses tâches de classification ensembliste.

Après avoir développé le concept de la classification multimodale et illustré ses performances et sa contribution indispensable pour divers systèmes actuels présentant un intérêt pratique en utilisant l'apprentissage d'ensemble. Le chapitre qui suit traite de la première contribution de cette thèse qui consiste à appliquer et à analyser l'impact de la multimodalité pour divers systèmes d'aide à la décision par l'intermédiaire de l'apprentissage d'ensemble utilisant l'apprentissage profond.

CHAPITRE 03

ANALYSE DE L'IMPACT DE LA MULTIMODALITE POUR L'AIDE A LA DÉCISION

Chapitre 03. Analyse de L'Impact de la Multimodalité pour L'Aide à la Décision

Ce chapitre porte sur l'application et l'analyse de l'impact de la *multimodalité* dans les systèmes d'aide à la décision au moyen de réseaux de neurones convolutionnels profonds « *CNNs* » et du classifieur *SVM*. La recherche actuelle implique des données d'entrée multimodales. En effet, les données traitées pour la conception d'un système d'aide à la décision peuvent être issues de différentes sources et natures. L'interaction entre les différents apprenants correspondant aux diverses modalités offre une opportunité de collecter des données riches et multimodales. Surmonter le défi de la collecte et de la compréhension de ces données présente le potentiel de fournir de nouveaux concepts permettant de soutenir les expériences d'apprentissage. Des exemples typiques de ces données multimodales incluent notamment des données d'images, du texte, etc. Dans notre étude, deux approches multimodales sont proposées afin d'analyser l'impact de la coopération de plusieurs modalités sur la qualité des systèmes d'aide à la décision en terme de capacité à générer les meilleures performances de décision. La première partie de cette étude consiste en une classification des données textuelles à travers un système multi-classifieurs basé sur la méthode de plongement lexical « *Word Embedding* ». La seconde vise à appliquer une nouvelle approche de fusion multimodale, appelée approche hybride dans le contexte de l'aide au diagnostic de la maladie du glaucome par le biais de la classification des images de la rétine.

1 Approche Multimodale Niveau Décision basée sur l'apprentissage profond Appliquée à la vérification des auteurs

1.1 Introduction

La classification des textes constitue une tâche de base et un élément clé dans de nombreuses applications de traitement automatique du langage naturel (*TALN ou NLP*) telles que le filtrage des informations, la recherche sur le web, l'analyse des sentiments et l'attribution d'auteur. Ainsi, de nombreuses études ont été menées par de nombreux chercheurs dans ce domaine.

L'attribution d'auteur est un sujet d'actualité traitant de la problématique de l'attribution d'un texte inconnu donné à un auteur, en tenant compte d'un ensemble d'auteurs candidats pour lesquels les échantillons de texte d'auteurs non contestés sont disponibles. Les échantillons de texte correspondant à tous les auteurs sont généralement appelés l'ensemble de référence, qui sera analysé afin d'obtenir le style d'écriture des auteurs candidats. La tâche d'attribution d'auteur est considérée comme un problème d'ensemble fermé si l'ensemble de référence contient l'auteur réel; sinon, il s'agit d'un problème d'ensemble ouvert. La vérification d'auteur est une tâche qui vise à savoir si un document inconnu donné a été écrit par un certain auteur (A) ou non. La figure 3.1 montre un exemple de processus de vérification d'auteur.

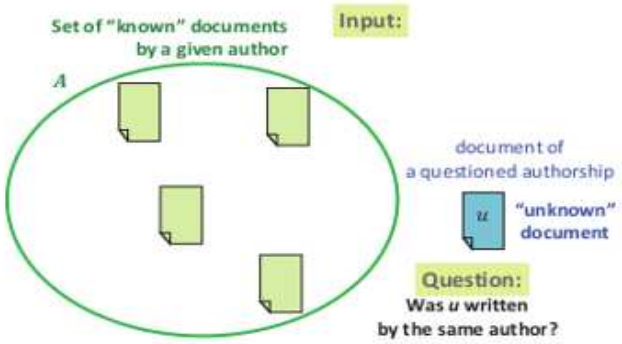

Figure 3.1 : Illustration du problème de vérification d'auteur [65]

L'ensemble de référence dans la vérification d'auteur comprend des échantillons du texte de l'auteur présumé. Il peut être analysé pour des caractéristiques bien définies qui montrent le style d'écriture de l'auteur. S'il y a une similitude dans le style d'écriture d'un texte donné, alors la paternité est vraie, sinon elle est fausse. Une série de problèmes de vérification d'auteur concernant les auteurs candidats donnés peut être considérée en tant que problème d'attribution d'auteur qui sera résolu si le problème de vérification est surmonté [66]. À l'inverse, la vérification d'auteur est une tâche particulière de l'attribution d'auteur auprès d'un ensemble ouvert d'auteurs candidats. À cet égard, la vérification d'auteur peut être vue comme un problème de classification binaire (0/1). Cela implique qu'un texte est classé en tant qu'appartenant à la catégorie donnée (attribué à l'auteur présumé) ou au contraire ne lui appartient pas (classé comme exemple négatif).

La représentation des caractéristiques est jugée comme un problème majeur dans le domaine de la classification des textes. Généralement, cette étape se focalise sur les modèles de sac de mots, d'unigrams, bigrams et n-grams pour l'extraction des caractéristiques. C'est pourquoi de multiples études ont été conduites dans ce domaine et de nombreuses techniques

d'apprentissage automatique ont été suggérées ces dernières années pour la tâche d'attribution/vérification d'auteur ces dernières années. Koppel et al. [67] ont présenté une méthode de vérification d'auteur appelée «démasquage» utilisant le classifieur machine à vecteurs de support (*SVM)* avec un noyau linéaire et ont traité la vérification d'auteur comme un problème de classification à classe unique (en ignorant les échantillons négatifs). Cette méthode a pour but de quantifier la dissemblance entre l'échantillon de document produit par le suspect et celui d'autres utilisateurs (c'est-à-dire des imposteurs). L'ensemble des caractéristiques est composé des 250 mots les plus fréquents. Bien que l'évaluation expérimentale de l'approche permette d'obtenir 95,70% de vérification correcte, les auteurs ont conclu que l'utilisation d'exemples négatifs pourrait améliorer les résultats. Iqbal et al. [68] ont mis à l'essai deux approches différentes. La première approche consiste à effectuer une vérification d'auteur en utilisant la classification; trois classifieurs sont étudiés, à savoir Adaboost.M1, Réseau Bayésien et Discriminative Multinomial Naive Bayes (DMNB). La deuxième approche réalise la vérification par régression; trois classifieurs sont adoptés, y compris la régression linéaire, SVM avec SMO et SVM utilisant le noyau RBF. L'ensemble des caractéristiques est constitué de 292 attributs, qui comprennent des attributs lexicaux, syntaxiques, idiosyncrasiques et spécifiques au contenu (mots clés couramment trouvés dans un domaine spécifique). Brocardo et al. [69] ont appliqué une technique de stylométrie pour la vérification d'auteur de courts messages en ligne, basée sur une combinaison d'apprentissage supervisé et d'analyse n-gram. Dans une autre étude de Brocardo et al. [70] dans laquelle un ensemble de caractéristiques a été utilisé, y compris des caractéristiques traditionnelles telles que des caractéristiques lexicales, syntaxiques, spécifiques à l'application, et de nouvelles caractéristiques extraites de l'analyse n-gram. L'approche fait appel à la stratégie de sélection des caractéristiques (le gain d'informations et les informations mutuelles) et au classifieur SVM pour la classification concernant la vérification d'auteur relative aux e-mails et aux tweets. Maitra et al. [71] ont utilisé au total 17 types de caractéristiques (ponctuation, longueur de phrase, vocabulaire, N-gramme, parties de discours) basées sur les mots et les styles en analysant la similarité entre les documents connus et leurs différences (ou similarités) par rapport à l'inconnu de l'ensemble de données *PAN at CLEF 2015*. Le classifieur *Random Forest (RF)* est adopté pour choisir les caractéristiques importantes et pour la phase de classification.

Récemment, le plongement lexical « *Word Embedding* » et les réseaux de neurones profonds ont fait l'objet d'un développement rapide, et ont permis de créer un nouveau domaine de

recherche dans plusieurs tâches du traitement automatique du langage naturel (*TALN*). Le plongement lexical est une catégorie d'approches d'apprentissage profond *« deep learning »* permettant de représenter les mots et les documents à l'aide d'une représentation vectorielle dense. Elle résout le problème de la limitation des données, qui peuvent conserver la syntaxe et la sémantique significatives figurant dans le texte [72]. Les réseaux de neurones convolutionels « *CNNs* » sont en mesure d'apprendre par eux-mêmes des caractéristiques utiles/pertinentes à partir de données brutes et de déterminer avec précision les phrases discriminantes dans un texte à l'aide d'une couche de *max-pooling*. Ainsi, les réseaux *CNNs* sont mieux à même de préserver la sémantique des textes par rapport à d'autres techniques d'apprentissage automatique. Ils ont démontré que cela représente une approche fructueuse pour la classification des textes [73]. De surcroît, en *TALN*, la position d'un mot dans la phrase est très déterminante. Le réseau de neurones récurrent (RNN) permet de simuler l'humain en lecture (de gauche à droite) mais le classifieur CNN peut lisser et ignorer l'ordre des mots. En fait, les RNNs, en particulier ceux qui utilisent les unités cachées de mémoire à long/court terme (*LSTM*) ont démontré d'excellentes performances en matière de représentation de texte, notamment pour les longues séquences de mots.

La présente étude suggère une approche multi-classifieurs pour la tâche de vérification d'auteur reposant sur la stratégie de plongement lexical « *Word Embedding* » pour produire la représentation vectorielle des mots en vue de la mettre à profit pour la phase de classification. Ainsi, trois classifieurs différents sont adoptés, à savoir les réseaux de neurones convolutionnels (*CNNs*), les CNNs récurrents (*RCNNs*) et les machines à vecteurs de support (*SVMs*) dans le but de faire intervenir différents points de vue de sorte que chaque classifieur/modalité fournisse au système une sorte d'informations complémentaires qui ne peuvent être obtenues à partir d'aucun des autres classifieurs pris individuellement. Une méthode de fusion de type classe est appliquée pour agréger les sorties des classifieurs afin de parvenir à la décision finale du système multimodal de classification.

Pour mieux décrire la démarche adoptée dans cette étude, une description du concept général du modèle *Word Embeddings (WE)* est introduite au début, suivie d'une explication détaillée des principales étapes de l'approche proposée ainsi que des résultats obtenus en terminant par une discussion et une conclusion.

1.2 L'apprentissage de la représentation vectorielle des mots : *word2vec*

Le plongement lexical (*WE*) est une méthode récente et très populaire dans le domaine du Traitement Automatique du Langage Naturel (*TALN)* permettant d'apprendre des représentations vectorielles de mots à partir de textes bruts tout en capturant les relations syntaxiques et sémantiques contrairement aux autres méthodes habituelles telles que la représentation *onehot*, les *unigrams*, *bigrams* et *n-grams*.

Word2vec est un outil couramment utilisé ces dernières années en raison de ses meilleurs résultats et de son apprentissage rapide. *Word2vec* consiste en un ensemble de modèles associés destinés à générer des plongements lexicaux. Il s'agit de réseaux de neurones artificiels peu profonds, à deux couches, formés pour reconstruire le contexte linguistique des mots. Cette technique dispose en entrée d'un grand corpus de textes et génère en sortie un espace vectoriel. Chaque mot du corpus sera donc représenté par un vecteur correspondant dans l'espace.

Nous pouvant distinguer deux architectures du *Word2vec*, à savoir le *sac de mots continu (CBOW)* ou le *skip-gram continu.* Le modèle *CBOW* se base sur le contexte pour prédire un mot cible et le modèle *skip-gram* procède à l'inverse de ce dernier; il utilise un mot pour prédire un contexte cible (voir la figure 3.2).

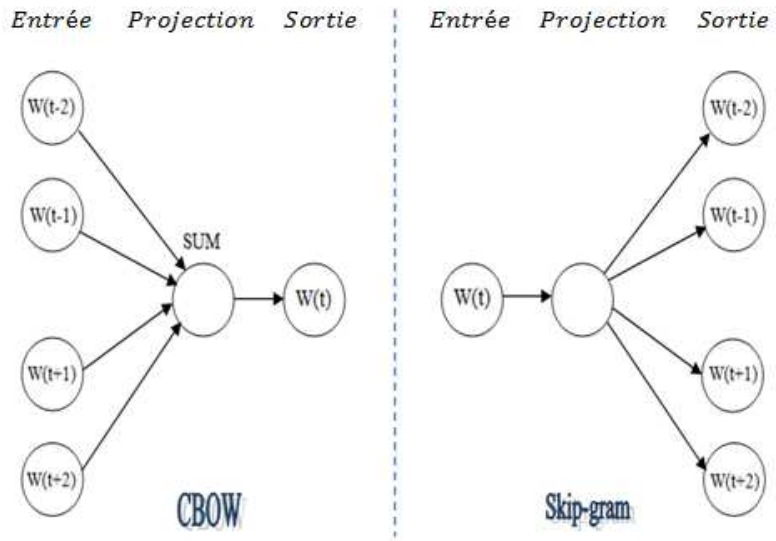

Figure 3.2 : Architectures de la méthode *Word2vec* [74]

La configuration de word2vec admet un ensemble important en termes d'hyper-paramètres. Des explications de certains paramètres sont présentées ci-dessous :

- *Layersize:* définit le nombre de caractéristiques dans le vecteur de mots. Ceci est égal au nombre de dimensions dans l'espace des caractéristiques.
- *Windowize:* indique la taille de la fenêtre contextuelle.
- *Minwordfrequency:* est le nombre minimum de fois qu'un mot doit apparaître dans le corpus.
- *Iterations:* il s'agit du nombre d'itérations (époques) effectuées pour chaque mini-lot pendant la phase d'apprentissage.
- *Learningrate:* correspond au taux d'apprentissage initial pour l'apprentissage du modèle.
- *Sampling:* détermine le recours ou non à un sous-échantillonnage.

1.3 L'approche proposée

La mise en place de système proposé, illustrée dans la figure 3.3, fut obtenue après plusieurs expériences et suite à une étude approfondie de la littérature concernant d'autres tâches de traitement automatique du langage naturel (*TALN)*.

1.3.1 Intégration de plusieurs modèles de classification

La combinaison de classifieurs a récemment été recommandée comme voie de recherche dans le but de rendre la reconnaissance plus fiable en utilisant la *complémentarité/diversité* qui peut exister entre les classifieurs; c'est la raison pour laquelle nous adoptons une approche multimodale qui se compose de divers classifieurs: les réseaux de neurones convolutionnels (*CNNs*), les réseaux de neurones convolutionnels récurrents (*RCNNs*) et les machines à vecteurs de support (*SVMs*). Les réseaux *CNN* et RNN sont deux architectures principales des réseaux de neurones profonds. Le classifieur *SVM* est considéré comme étant l'un des meilleurs classifieurs dans le cadre d'un problème de classification bi-classe tel que la problématique en question.

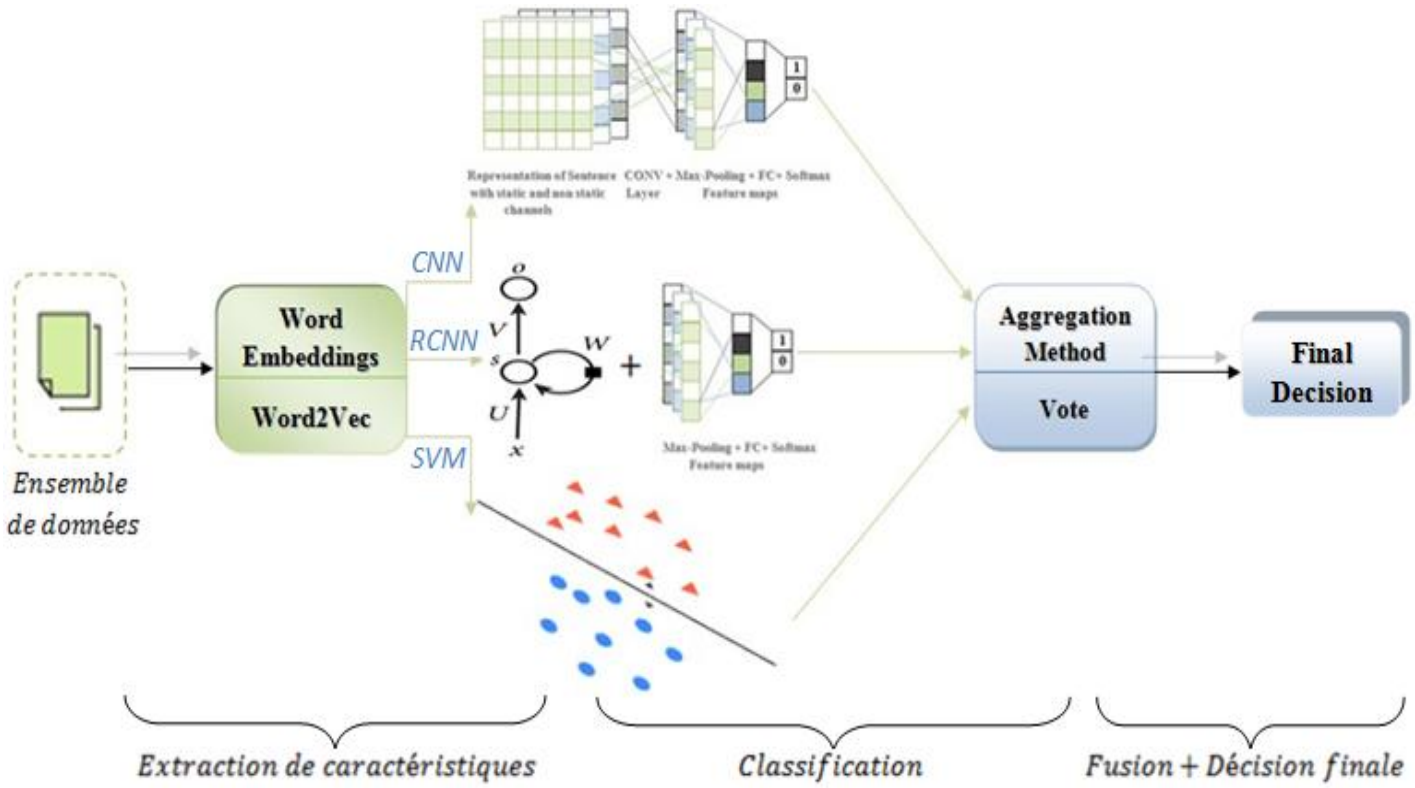

Figure 3.3: Architecture de l'approche multimodale proposée pour la vérification d'auteur [65]

Dans le cadre de cette expérience, le réseau *CNN* est employé à la fois comme extracteur de caractéristiques et un classifieur, qui est capable d'extraire des caractéristiques invariantes de position. Simultanément, le réseau *RNN* à mémoire court-terme et longue-terme (*LSTM*) est utilisé, de manière à appliquer une structure récurrente pour capturer autant que possible les informations contextuelles lors de l'apprentissage des représentations des mots, ce qui peut introduire considérablement moins de bruit par rapport aux réseaux de neurones traditionnels. En outre, une couche de *MAX-pooling* est mise en place, déterminant ainsi de façon automatique les caractéristiques qui jouent un rôle important dans la classification des textes. Le dernier élément utilisé dans le processus de classification est le classifieur *SVM* ayant comme donnée d'entrée la représentation vectorielle des mots « *features map* » et joue le rôle du classifieur ou une des modalités adoptées.

La décision finale de l'approche proposée est obtenue grâce à la mise en œuvre d'une technique de fusion basée sur le vote. Le *vote à la majorité* est jugé être l'une des méthodes de fusion (de type classe) les plus efficaces et simple. Dans le cadre du vote à la majorité, chaque classifieur prend une décision concernant l'étiquette d'un échantillon d'entrée. La classe ayant obtenu le plus grand nombre de votes est désignée comme la classe représentative de tous les

classifieurs de l'ensemble. Soit $C= [c_1, c_2, \dots c_L]$ un ensemble de L classifieurs, x est l'échantillon d'entrée et $c_{i,j}$ représente la sortie du i^{th} classifieur pour la j^{th} classe. La décision finale utilisant le vote à la majorité (VM) peut être définie comme suit :

$$MV(x) = max_{j\in\Omega} \sum_{i=1}^{L} c_{i,j} \tag{3.1}$$

Cependant, la précision spécifique de chaque classifieur n'est pas prise en compte dans la décision finale. Cette dernière étant considérée comme le principal inconvénient des méthodes de vote à la majorité. En général, les classifieurs sélectionnés ne possèdent pas une compétence similaire. Ainsi, la méthode de vote pondéré est employée pour combiner la décision des classifieurs choisis [75]. Dans cette méthode d'agrégation, le résultat de chaque classifieur est pondéré par un coefficient qui influence le processus de combinaison. À noter que w_i correspond au poids du i^{th} classifieur, le vote à la majorité pondérée (VMP) est ainsi défini :

$$VMP(x) = max_{j\in\Omega} \sum_{i=1}^{L} w_i c_{i,j} \tag{3.2}$$

and $\sum_{i=1}^{L} w_i = 1$

De nombreux schémas ont été suggérés pour estimer le poids des classifieurs [65,76]. Habituellement, ces poids sont calculés en utilisant la précision spécifique de chaque classifieur. Soit a_i et a_j les précisions des i^{th} et j^{th} classifieurs sur l'ensemble de validation. Le poids w_i est alors calculé par :

$$w_i = \frac{a_i}{\sum_j a_j} \tag{3.3}$$

Dans cette étude, le schéma de *vote pondéré meilleur-pire* ($VPMP$) ou *Best-Worst Weighted Vote* ($BWWV$) [75] est appliqué à titre de mesure pour quantifier les poids. L'idée principale derrière ce schéma est d'identifier les meilleurs et les pires membres de l'ensemble en utilisant leur erreur estimée sur l'ensemble de validation; les valeurs a_i sont déterminées à l'aide de l'expression suivante:

$$a_i = \mathbf{1} - \frac{e_k - e_b}{e_w - e_b} \tag{3.4}$$

où e_w indique l'erreur maximale parmi les classifieurs, et e_b est l'erreur minimale. La valeur de a_i varie entre $[0,1]$, dans laquelle la valeur 0 désigne le pire classifieur et la valeur 1 correspond au meilleur classifieur.

1.3.2 Architecture et conception du modèle CNN

En ce qui concerne l'architecture du modèle de CNN proposé, quatre (04) couches sont utilisées dans cette étude qui sont structurées comme suit : une couche d'entrée E, une couche de convolution C, une couche de pooling P, une couche dense FC et une couche de sortie S.

Comme déjà discuté (dans le premier chapitre) à propos du principe du réseau *CNN*, les CNNs sont actuellement les modèles les plus puissants pour diverses tâches de classification, y compris la classification d'images et de texte [12, 77]. À l'entrée d'un réseau CNN, une image se présente sous la forme d'une matrice de pixels. Le fonctionnement de base de ce réseau profond consiste à utiliser la convolution des images et des filtres pour engendrer des caractéristiques invariables qui sont ensuite transférées à la couche suivante (*max-pooling*). Les caractéristiques de la couche pooling sont sous-échantillonnées au moyen de divers filtres afin de générer des caractéristiques plus invariantes et abstraites, et le processus se continue jusqu'à ce que le résultat final soit obtenu.

Les textes, comme les images, une matrice représentant les phrases et les mots de notre corpus sont substitués aux pixels d'une image; chaque ligne de la matrice correspond à un mot. Il est maintenant possible d'appliquer des filtres de la couche de convolution et d'utiliser de manière préférentielle des filtres de largeur équivalente à la dimension des vecteurs [77]. Il est à noter que $s \times d$ correspond à la dimension de la matrice de la phrase, où s représente la taille de la phrase et d est la dimension d'un vecteur de mots.

La couche de *pooling* ou sous-échantillonnage présente plusieurs avantages, notamment de garantir que la matrice de sortie aura une taille fixe égale au nombre de filtres utilisés (pouvant ensuite être introduit dans une couche *softmax*). Ainsi, elle permet d'obtenir toujours les mêmes dimensions de sortie en utilisant des phrases de différentes longueurs et des filtres de différentes tailles de régions. En outre, cette méthode garantit également le processus de sélection des caractéristiques dans la mesure où chaque filtre tentera de détecter une caractéristique spécifique (seule la valeur maximale de chaque carte de caractéristiques est prise en compte) afin de générer les caractéristiques pertinentes pour effectuer la classification.

1.3.3 Configuration du modèle CNN

La mise en œuvre de notre réseau profond CNN est effectuée après plusieurs tests de performance. Nous débutons par la couche de convolution ayant comme entrée le plongement de mots accompagné d'une fonction d'activation « *Tanh* ». Après chaque couche de convolution, une normalisation des lots est réalisée. Comme méthode de régularisation du poids et du biais, nous utilisons une *descente de gradient stochastique* avec un *momentum* de $0{,}9$, $L2$ $(0{,}0005)$, et un faible taux d'apprentissage (*learning rate*) de $0{,}0001$ pour entraîner notre réseau. Les couches de convolution et de pooling sont structurées l'une après l'autre à l'aide d'une fonction « *ReLU* » largement utilisée [8, 12]. Un pas de $(1{,}1)$ est effectué dans le processus de filtrage pour les deux couches, ainsi qu'une taille de noyau 2×2. La fonction *Erreur quadratique moyenne* (*MSE*) est adoptée pour optimiser la fonction de perte. Enfin, la fonction *Softmax* est utilisée pour la classification.

1.4 Résultats expérimentaux

La présente étude propose une nouvelle approche multi-classifieurs pour la vérification d'auteur basée sur la technique de plongement de mots *word2vec*. Dans le but de tirer profit de la complémentarité par fusion de plusieurs classifieurs d'une part, et d'analyser le comportement de l'approche ensembliste multimodale élaborée selon le modèle word2vec adopté (*Skip-Gram* ou *CBOW*) d'autre part, trois modèles de classification sont utilisés, à savoir les classifieurs *CNN*, *RCNN* et *SVM*. La mise en œuvre du système proposé est réalisée via la plateforme *Deeplearning4j*[2]. Deeplearning4j est une bibliothèque d'apprentissage profond distribuée, open source, écrite pour *Java* et *Scala*. Dans les sections qui suivent, les détails des expériences et leurs résultats sont présentés.

1.4.1 Description de l'ensemble de données utilisé

Dans le cadre de cette expérience, les ensembles de données de *PAN at CLEF 2015*[3] sont utilisés. La base de données a été développée, écrite en anglais, comprenant 100 auteurs; pour chaque auteur A_i, un document connu et un autre inconnu sont fournis et la tâche consiste à déterminer si un document connu et un document en question appartient au même auteur ou pas. Les données de vérité du corpus d'apprentissage sont fondées sur le fichier *truth.txt*

[2] https://deeplearning4j.org/
[3] http://pan.webis.de/data.html

comprenant une ligne pour chaque problème, la bonne réponse binaire (Y signifie que les documents connus et interrogés sont du même auteur ; N signifie le contraire).

1.4.2 Le plongement de mots à l'aide de l'approche Word2Vec

Dans le but de construire le modèle de représentation vectorielle des mots (*Word Embeddings*) approprié, diverses expériences ont été menées concernant la dimensionnalité vectorielle des mots et les différentes architectures *word2vec*. Les vecteurs de mots résultants sont utilisés en vue de former les différents classifieurs adoptés pour la classification des textes. Le tableau 1 illustre les paramètres d'apprentissage utilisés pour plusieurs modèles basés sur *Skip-Gram* et *CBOW*.

Tableau 3.1: Configuration des modèles *Skip-Gram* et *CBOW*

Modèle	Layersize	Windowsize	Minwordfrequency	Iterations	Learningrate	Sampling
Skip-Gram	200	10	10	3	1.0E-4	1.0E-5
CBOW	200	10	10	3	1.0E-4	1.0E-5

En effet, les résultats fournis démontrent que la méthode *Skip-Gram* génère de meilleures performances et des données/réponses plus précises en ce qui concerne le corpus utilisé.

1.4.3 Les critères d'évaluation

Afin d'évaluer les performances de notre système, les principales métriques d'évaluation utilisées sont: $la\ précision\ (accuracy)$, $la\ sensibilité$ et $la\ spécificité$ qui sont définies comme suit:

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{3.5}$$

$$\text{Sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.6}$$

$$\text{Spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{3.7}$$

Où, Vrai Positif (VP) indique le nombre de documents interrogés (*connus*) avec un test positif (*connus*), Vrai Négatif (VN) désigne le nombre de documents interrogés (*inconnus*) ayant un test négatif (*inconnus*), Faux Positif (FP) est le nombre de documents interrogés (*inconnus*)

avoir un test positif (*connus*) et Faux Négatif (FN) représente le nombre de documents interrogés (*connus*) dont le test est négatif (*inconnus*).

1.4.4 Résultats et discussions

Afin d'aboutir à une décision unique optimale de l'approche ensembliste suggérée, les performances individuelles de chaque classifieur sont récapitulées dans les tableaux suivants selon les deux modèles *skip-gram* et *CBOW*, en analysant les différentes techniques de combinaison adoptées (*VM*, *VMP* et *VPMP*).

Comme indiqué dans le tableau 2 (*CNN*), 48 et 47 documents considérés comme des documents connus sont exactement classés comme des documents connus respectivement selon les techniques *Skip-Gram* et *CBOW*, 47 et 44 documents inconnus sont correctement classés comme des documents inconnus respectivement en fonction des techniques *Skip-Gram* et *CBOW*.

Tableau 3.2: Matrice de confusion du modèle CNN

Méthode ***Word2vec*** utilisée	Document	**Interrogé 'Y'**	**Interrogé 'N'**
Skip-Gram	**Connu 'Y'**	**48**	2
	Inconnu 'N'	3	**47**
CBOW	**Connu 'Y'**	**47**	3
	Inconnu 'N'	6	**44**

En résumé, 95 documents ont été correctement étiquetés via le modèle *CNN* proposé, soit une précision de 95% en utilisant la technique Skip-Gram dont la sensibilité est de 96% et la spécificité de 94%. Dans le cas de la technique CBOW, 91 documents ont été précisément identifiés, ce qui donne une précision de 91%, avec une sensibilité de 94% et une spécificité de 88%.

Le tableau 3 (*RCNN*) illustre que, 48 et 45 documents connus sont formellement classés comme des documents connus respectivement en fonction les techniques *Skip-Gram* et *CBOW*, 49 et 48 documents inconnus sont correctement classés comme des documents inconnus respectivement selon les techniques *Skip-Gram* et *CBOW*.

Tableau 3.3: Matrice de confusion du modèle RCNN

Méthode *Word2vec* utilisée	Document	Interrogé 'Y'	Interrogé 'N'
Skip-Gram	**Connu 'Y'**	**48**	2
	Inconnu 'N'	1	**49**
CBOW	**Connu 'Y'**	**45**	5
	Inconnu 'N'	2	**48**

En résumant, 97 documents ont été correctement identifiés via le modèle *RCNN* suggéré, ce qui donne une précision de 97% en utilisant la technique *Skip-Gram* avec une sensibilité de 96%, une spécificité de 98%. Concernant la technique *CBOW*, 93 documents ont été correctement étiquetés, ce qui représente une précision de 93% avec une sensibilité de 90% et une spécificité de 96%.

Le tableau 4 (*SVM*) montre que 47 et 44 documents connus sont correctement classés comme des documents connus respectivement utilisant les techniques *Skip-Gram* et *CBOW*, 48 documents inconnus sont identifiés exactement comme des documents inconnus respectivement selon les deux techniques *Skip-Gram* et *CBOW*.

Tableau 3.4: Matrice de confusion du modèle SVM

Méthode *Word2vec* utilisée	Document	Interrogé 'Y'	Interrogé 'N'
Skip-Gram	**Connu 'Y'**	**47**	3
	Inconnu 'N'	2	**48**
CBOW	**Connu 'Y'**	**44**	6
	Inconnu 'N'	2	**48**

En général, 95 documents ont été correctement identifiés au moyen du modèle *SVM* proposé, ce qui se traduit par une précision de 95% en appliquant la technique *Skip-Gram* pour une sensibilité de 94% et une spécificité de 96%. En utilisant la technique *CBOW*, 92 documents ont été correctement étiquetés, ce qui donne une précision de 92% avec une sensibilité de 88% et une spécificité de 96%.

En comparant les différents résultats ainsi obtenus, on peut en déduire que la technique *Skip-Gram* est plus performante que celle de *CBOW* en ce qui concerne les trois classifieurs utilisés

(*CNN*, *RCNN* et *SVM*) et la surpasse dans tous les cas de figure. Le tableau suivant récapitule les résultats obtenus précédemment.

Tableau 3.5: Récapitulatif des résultats obtenus de tous les classifieurs utilisés

Modèle	Précision	Sensibilité	Spécificité
CNN+Skip-gram	95%	96%	94%
CNN+CBOW	91%	94%	88%
RCNN+Skip-gram	97%	96%	98%
RCNN+CBOW	93%	90%	96%
SVM+ Skip-gram	95%	94%	96%
SVM+CBOW	92%	88%	96%

La sensibilité globale de notre approche est améliorée par le biais de la combinaison des résultats des trois modèles, à savoir *CNN*, *RCNN* et *SVM* sur la base de la méthode *Skip-Gram* en vue de parvenir à une décision finale unique concernant la tâche de vérification d'auteur en utilisant les trois schémas d'agrégation de type classe: le vote à la majorité (VM), le vote à la majorité pondérée (VMP) et le vote pondéré meilleur-pire ($VPMP$). Les résultats obtenus dans le cadre de cette combinaison sont illustrés dans le tableau (6) ci-dessous.

Tableau 3.6: Résultats obtenus à travers les trois techniques de fusion utilisées

Méthode utilisée	Précision	Sensibilité	Spécificité
VM	95%	94%	96%
VMP	97%	96%	98%
VPMP	98%	98%	98%

En effet, il ressort de l'analyse des résultats en matière d'expérimentation que la combinaison des classifieurs est meilleure que leur apprentissage indépendant. Par ailleurs, la règle $VPMP$ est plus appropriée par rapport à d'autres techniques (VM et VMP).

1.5 Conclusion

La vérification d'auteur est une tâche importante et représente un domaine de recherche très récent dans le contexte des applications *TALN*, qui vise à déterminer pour un document inconnu donné s'il appartient ou non à un certain auteur.

Le présent travail a pour objectif l'analyse de la coopération de plusieurs modalités de type décision ou prédicteurs a garantissant la robustesse du système d'aide à la décision en se basant sur des données textuelles.

Les résultats expérimentaux prouvent que la classification ensembliste est plus avantageuse par rapport à l'apprentissage séparé des classifieurs ou l'apprentissage monomodal tout en bénéficiant de la *complémentarité* qui peut exister entre les classifieurs. En outre, la règle d'agrégation du *vote pondéré meilleur-pire* ($VPMP$) est plus efficace que le vote à la majorité (VM) et le vote à la majorité pondérée (VMP) en appliquant la méthode *Skip-Gram*.

En effet, cette approche traite le concept Multimodalité juste au niveau décision (représenté dans notre cas par les classifieurs) qui peut ignorer certaines informations utiles/pertinentes pour la classification qui sont disponibles dans les niveaux inférieurs (niveau de caractéristiques).

L'étude suivante consiste à revoir la richesse apportée par l'utilisation de la *multimodalité* en l'intégrant à différents niveaux à la fois (deux niveaux ont été adoptés qui sont le niveau caractéristiques ou espace de représentation et le niveau décision) en proposant une nouvelle approche hybride.

2 Approche Multimodale Basée Caractéristiques avec Fusion des Classifieurs pour le Diagnostic du Glaucome

2.1 Introduction

La maladie du glaucome est considérée comme la seconde cause de la détérioration visuelle après la dégénérescence maculaire liée à l'âge (DMA). Le nombre de personnes touchées dans le monde dépasserait les 70 millions d'ici 2020 [78]. Le glaucome peut être pris en charge, mais peut également provoquer la cécité s'il n'est pas détecté à temps. Le glaucome est une maladie oculaire qui touche principalement les personnes de plus de 45 ans. Cette maladie peut causer une lésion du nerf optique ; le nerf débute avec la rétine à l'arrière de l'œil et transporte les images au cerveau. Lorsque ce nerf est endommagé, le champ visuel est réduit, la vision est alors modifiée et cela peut entraîner une cécité à long terme. Dans la plupart des cas, le glaucome est lié à une augmentation de la pression à l'intérieur de l'œil, également appelée hypertension intraoculaire ou pression intraoculaire (PIO).

La pression intraoculaire est généralement mesurée par un test de tonométrie, qui est un test élémentaire, puisqu'une PIO élevée constitue un facteur de risque important de l'apparition du glaucome. Cependant, une PIO élevée n'est pas toujours synonyme de glaucome, et une PIO normale ne signifie pas nécessairement qu'un patient ne sera jamais atteint de glaucome. Le glaucome est une maladie délicate, de sorte qu'il est très difficile pour une personne de remarquer une éventuelle déficience visuelle liée au glaucome en raison de l'absence totale de symptômes.

Le glaucome, lorsqu'il est diagnostiqué à temps, peut être traité et la vision peut ensuite être stabilisée. Par conséquent, s'il n'est pas détecté et n'est pas pris en charge à temps (précocement), le glaucome peut évoluer et entraîner une cécité complète. En revanche, la réduction du taux de prise en charge de cette maladie silencieuse et grave fait partie des principaux intérêts de la santé publique, afin de la prendre en charge dès sa première apparition et de contrôler sa progression, permettant ainsi un meilleur diagnostic du glaucome.

Dans le but d'aider les ophtalmologistes à dépister le glaucome à un stade précoce, diverses études ont été orientées vers les systèmes d'aide à la décision (SAD) en matière de glaucome. L'objectif principal des systèmes automatisés réside dans l'amélioration de la précision du

diagnostic. En fait, ils sont souvent utilisés comme un deuxième avis par les médecins pour parvenir au diagnostic final [79], ce qui permet de réduire les erreurs humaines, afin d'offrir un dépistage uniforme à grande échelle et à un meilleur prix.

Une fois formés, les ordinateurs/systèmes peuvent obtenir des classifications beaucoup plus rapides, ce qui aide les médecins dans leur classification en temps réel. La classification du glaucome a subi un excellent développement ces dernières années, notamment en ce qui concerne l'utilisation des paradigmes d'apprentissage automatique. En effet, Dans le domaine de l'apprentissage automatique et du diagnostic des maladies, les caractéristiques sont considérées comme étant les informations les plus importantes dans la reconnaissance des formes.

Ces dernières années, les systèmes de classification ont fait appel à des techniques d'extraction de caractéristiques telles que les primitives de forme et de texture, notamment celles de la matrice de cooccurrence de niveau de gris d'Haralick [80, 81], de dimension fractale [82], de spectre d'ordre supérieur [83, 84], des caractéristiques basées sur les ondelettes [84, 85, 86], des caractéristiques du modèle de configuration local [87], des caractéristiques de correntropie [86], des caractéristiques de transformée de fourier rapide [88] et de descripteur de caractéristiques *GIST* [89]. La plupart de ces travaux [80, 82, 84, 88, 89] ont utilisé la machine à vecteurs de support (*SVM*) comme technique de classification. Maheshwari et al. [86] ont utilisé une variante du classifieur SVM à savoir le classifieur SVM des moindres carrés (*LS-SVM*) avec une fonction de base radiale (*RBF*), une ondelette de Morlet et des noyaux d'ondelettes à chapeau mexicain. La précision de l'approche proposée est de 98,33% en utilisant la méthode de validation croisée à $k = 3$ blocs. Acharya et al. [87] ont analysé plusieurs classifieurs dans le cadre du diagnostic du glaucome, à savoir le réseau de neurones probabiliste (*PNN*), l'arbre de décision (*DT*), les k plus proches voisins (*k-ppv*), l'algorithme *SVM* et l'analyse discriminante linéaire (*ADL*), dont le classifieur K-ppv a permis d'obtenir un meilleur taux de précision de 95,7%. Kumbhare et al. [90] se sont basés sur les classifieurs de Naïve Bayes (*NB*) et de distance minimale pour le diagnostic automatique du glaucome en utilisant les caractéristiques de spectre d'ordre supérieur (HOS) et de texture (matrice de cooccurrence au niveau du gris et matrice à longueur de plage) avec un taux de classification de 91%. Singh et al. [91] ont proposé un système de diagnostic du glaucome faisant appel à cinq des principaux algorithmes d'apprentissage automatique, Random Forest (*RF*), *NB*, *k-ppv*, *SVM* et réseau de neurones artificiels (*RNA*), de manière à choisir le

classifieur le plus performant. L'approche suggérée utilise la sélection d'attributs évolutifs en ce qui concerne la sélection des caractéristiques et la méthode d'analyse en composantes principales (*ACP*) comme technique de réduction de ces dernières en recourant aux caractéristiques d'ondelettes du disque optique segmenté. Les résultats obtenus indiquent un taux de précision de 94,7%. Maheshwari et al. [92] ont fait recours au classifieur *LS-SVM* dans le but de diagnostiquer de manière automatique la maladie du glaucome. L'algorithme de *Relief Binaire* a également été utilisé en vue d'extraire les caractéristiques pertinentes. Le taux de reconnaissance obtenu était de 95,19% en appliquant la méthode de validation croisée à $k = 3$ blocs. Kausu et al. [93] ont suggéré la mise en place d'une méthode de détection automatique du glaucome utilisant les caractéristiques des ondelettes et les caractéristiques morphologiques à partir des images du fond d'œil; basée sur la segmentation de la région d'intérêt (*ROI*). La précision de 97,67% est obtenue en employant le classifieur *MLP* au moyen de la technique de validation croisée à $k = 10$ blocs.

De tels résultats sont encourageants [80-94], mais restent insuffisants dans la mesure où la représentation optimale selon la base utilisée n'est pas assurée, et que l'on ne sait pas si les caractéristiques artisanales sont optimales dans leurs performances. Les méthodes classiques de classification reposant sur la forme utilisent également des techniques d'extraction de caractéristiques afin de représenter la forme - tout en testant/analysant diverses familles conformément à la base de données appliquée-. Le choix des caractéristiques n'est pas justifiable et ne garantit en aucun cas la capacité des caractéristiques retenues à représenter de nouvelles images ; en effet, la modification de la base initiale ou son enrichissement remet en cause les caractéristiques déjà adaptées et implique impérativement de refaire la phase d'extraction des caractéristiques.

Récemment, le réseau de neurones convolutionnels (*CNN*) s'est généralisé et constitue une architecture d'apprentissage profond qui génère des caractéristiques de manière automatique [95]. En d'autres termes, CNN permet d'apprendre et d'extraire les caractéristiques les plus discriminantes des données en cours d'apprentissage. Il a été démontré que cette architecture donne des résultats statistiquement impressionnants dans le cadre d'applications de reconnaissance d'images [8, 12, 96]. De nombreuses études ont révélé des performances supérieures en utilisant des réseaux de neurones convolutionnels profonds (*CNNs*) dans le contexte du diagnostic de la maladie du glaucome. Chen et al. [97] se sont basés sur le réseau de neurones convolutionnels (*CNN*) pour la détection du glaucome au moyen de six couches.

Le réseau *CNN* est formé à travers des images segmentées de la région d'intérêt (*ROI*) en utilisant 1 676 images issues de l'ensemble des données *SCES* et 650 images provenant de l'ensemble des données *ORIGA* afin de valider l'efficacité de l'approche proposée. Ces résultats reposent sur des valeurs de l'aire sous la courbe ROC «AUC» représentant respectivement 83,1% et 88,7% des ensembles de données *ORIGA* et *SCES*. Chai et al. [98] ont mis au point un système de classification pour le glaucome utilisant un réseau *CNN* à deux branches afin de pouvoir analyser à la fois l'image entière et la région du disque optique extraite automatiquement à l'aide du modèle *Faster-RCNN*. Les auteurs ont intégré une couche de fusion permettant de combiner les caractéristiques extraites à partir de deux branches et une couche entièrement connectée en ce qui concerne la phase-classification. Le taux de classification optimal étant de 81,69% en utilisant cinq couches convolutionnelles. Zilly et al. [99] ont eu recours à la technique de l'apprentissage d'ensemble dans le but de segmenter la cupule et le disque optiques à partir des images de la rétine en utilisant également une architecture *CNN*, tout en calculant le rapport cupule/disque pour la classification automatique du glaucome. Afin de diminuer la complexité des calculs et de fournir de meilleures performances, une technique d'échantillonnage entropique est adoptée. Orlando et al. [100] ont élaboré un modèle *CNN* permettant la classification automatique du glaucome en utilisant deux architectures distinctes, à savoir *OverFeat* et *VGG-S*, à partir des images du fond d'œil. La segmentation de la zone de la tête du nerf optique (*ONH*) et la technique d'incrustation des vaisseaux ont été appliquées aux images du fond d'œil pour améliorer la qualité de l'image et évaluer ainsi le perfectionnement de la discrimination des caractéristiques. La performance de cette méthode est estimée en fonction de l'aire sous la courbe ROC « AUC»; les valeurs *AUC* des deux architectures CNNs (OverFeat et VGG-S) sont respectivement de 76,3% et 71,8%. Raghavendra et al. [101] ont proposé un système de diagnostic assisté par ordinateur pour la classification automatique du glaucome en utilisant le réseau *CNN* et le classifieur *ADL*. Au cours de cette étude, les auteurs se sont servis de 1 426 images du fond d'oeil afin de former-*CNN* en employant dix-huit couches et en obtenant de meilleures performances diagnostiques.

En réalité, le principal avantage de l'utilisation du réseau de neurones convolutionnels profonds (*CNN*) est de prendre l'image entière au lieu de la partie défectueuse, permettant ainsi d'éviter la conception sophistiquée de caractéristiques artisanales, ce qui représente une démarche fastidieuse; cela permet un gain de temps et de mémoire important. Ainsi, en tant que second avantage de la mise en œuvre du *CNN*, il permet également de ne pas avoir

recours à la segmentation et fournit des caractéristiques efficaces en vue de classer de manière appropriée les patients malades et les patients non malades [8, 12, 96]. Par ailleurs, le réseau *CNN* ne fait intervenir aucune étape de prétraitement susceptible d'affecter les performances.

Néanmoins, compte tenu du problème majeur de la limite des données et du bon choix des hyper-paramètres, tels que le nombre de filtres, leur noyau et la forme de *max-pooling*, etc., il est difficile de savoir si les caractéristiques considérées via le réseau *CNN* sont vraiment les plus représentatives. Par conséquent, il serait préférable de les fusionner avec d'autres familles de caractéristiques dans le but d'obtenir de meilleures performances.

En règle générale, les défis de la classification se résument à trouver une meilleure zone de décision qui sépare les objets en catégories/classes. Afin de définir la meilleure séparation, nous introduisons le concept de marge ou de plan entre deux catégories. Pour rendre l'idée plus simple, nous montrons un exemple dans un espace bidimensionnel afin d'expliquer comment chaque classifieur détermine sa propre marge séparant deux classes. Les points rouges représentent les échantillons de la première classe et les points bleus ceux de la seconde. Cette idée peut être généralisée à un espace de grandes dimensions. L'idée est illustrée dans la figure 3.4.

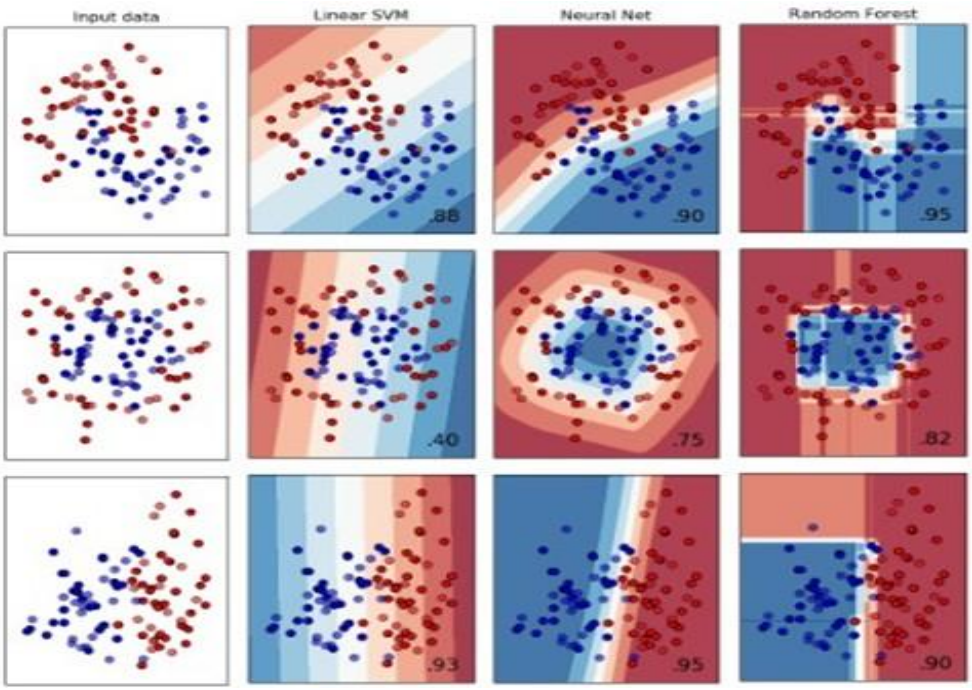

Figure 3.4: Différentes surfaces de séparation générées par divers classifieurs [8]

Il convient de souligner que chaque classifieur possède sa propre façon de générer la marge qui sépare les classes et le modèle résultant diffère d'un classifieur d'un autre; en effet, il est

généralement impossible de construire une partition parfaite de l'espace, de sorte que le rôle du classifieur sera souvent de donner une probabilité d'appartenance d'un objet à une classe. Par conséquent, le choix du classifieur n'est pas évident dans la mesure où on ne dispose pas d'un seul classifieur considéré comme le meilleur pour résoudre tous les problèmes. Ce choix est conditionné par le contenu de la base de données et la dispersion des données.

Récemment, la combinaison de classifieurs a fait l'objet d'une proposition de voie de recherche permettant une reconnaissance plus fiable en utilisant la complémentarité qui peut exister entre les classifieurs. Alors que les premières expériences de combinaison de classifieurs remontent aux années 1980 [102], cette technique est devenue un moyen de plus en plus utilisé en vue d'améliorer la qualité des systèmes de reconnaissance et ce, dans plusieurs applications, à savoir la reconnaissance d'images médicales [103], la reconnaissance de chiffres manuscrits [104], la reconnaissance de visages [105] et la reconnaissance vocale [106]; ces systèmes diffèrent selon le type de sortie des classifieurs combinés et par la nature des classifieurs utilisés.

La principale raison qui motive le véritable intérêt que porte la communauté de la reconnaissance des formes à la combinaison de classifieurs réside dans sa capacité à prendre en compte un nombre de caractéristiques important utilisées par différents classifieurs en exploitant les performances et le comportement marginaux de chacun de ces classifieurs.

Sur le plan de la reconnaissance des formes, et plus particulièrement dans le contexte du diagnostic médical à travers le contenu des images, il est avantageux de prendre en compte plusieurs modalités lors de la prise de décision. En fait, la *multimodalité* est en mesure de fournir des informations exhaustives à propos du contenu des images, d'augmenter les capacités d'interprétation, de perfectionner les caractéristiques d'analyse et de produire des résultats plus fiables. Toutefois, la majorité des recherches existantes se concentrent désormais sur une seule modalité afin de parvenir à diagnostiquer une maladie telle que le glaucome [107], en dépit d'études récentes qui ont montré que l'apprentissage à l'aide de données multimodales peut fournir des informations complémentaires [108] dans le but d'obtenir des performances accrues [109] en matière d'extraction de caractéristiques et de classification.

La présente contribution expose une nouvelle méthode de classification multimodale basée essentiellement sur deux types de classifieurs communément répandus, à savoir le réseau de

neurones convolutionnels profonds (*CNN*) et la machine à vecteurs de support (*SVM*), en utilisant des données multimodales (*au niveau de ressource*) et des caractéristiques multimodales aux fins du diagnostic du glaucome à partir d'images du fond de la rétine provenant de l'ensemble de données *RIM-ONE* dans le but de détecter de manière appropriée le nerf optique et d'obtenir ainsi un meilleur taux de classification permettant un diagnostic plus fiable du glaucome.

Dans ce contexte, notre travail constitue le premier à proposer deux niveaux de *multimodalité* en vue de la classification du glaucome :

- ✓ Cette étude s'appuie sur deux modalités relatives aux images d'entrée, soit les images couleur *RVB* originales et une autre modalité binaire utilisant la technique dite *d'Otsu* dans laquelle chaque modalité apporte au système certains types d'informations ne pouvant être déduites ou obtenues à travers d'autres modalités. En termes mathématiques, ces informations ajoutées sont désignées sous le nom de *diversité*.
- ✓ Le recours à deux modalités de représentation des images (RVB et binaire), correspondant aux caractéristiques générées automatiquement au moyen de CNN et aux caractéristiques artisanales de texture et de forme (exploitées par le classifieur *SVM*); ce travail vise à tirer profit de ces représentations en faisant appel aux techniques d'apprentissage d'ensemble. En effet, chaque représentation engendre sa propre vision de l'image dont la combinaison de plusieurs points de vue augmente indéniablement la performance du diagnostic.
- ✓ De même, l'ajout d'un cinquième modèle qui permet de combiner les caractéristiques extraites grâce aux couches de convolution des deux modalités d'image utilisant le classifieur *SVM* en ce qui concerne la phase de classification.
- ✓ Ce travail propose une nouvelle approche de fusion appelée *fusion hybride*; elle comprend la mise en concaténation des caractéristiques de différentes modalités (fusion précoce) et la classification multimodale (fusion tardive) en utilisant la technique de vote pondéré meilleur-pire (VPMP) ou Best-Worst Weighted Vote (BWWV) [75] afin de générer la décision finale de notre système multimodal.

Le complément de cette étude illustre en premier lieu le schéma proposé de fusion hybride en présentant le concept de base des techniques de fusion précoce et tardive, ensuite une description approfondie des grandes lignes de notre architecture CNN-SVM suggérée pour le

diagnostic du glaucome ainsi que les résultats expérimentaux de ce travail, se terminant par une conclusion.

2.2 Fusion multimodale : *vers une stratégie de fusion hybride*

Comme déjà développé dans le deuxième chapitre au sujet du nouveau concept de *multimodalité*, ce terme désigne la mise en œuvre de plusieurs modalités en vue d'accomplir en même temps une tâche donnée. En effet, dans la pratique, de nombreuses applications impliquent un traitement de données multimodales. Compte tenu de la richesse des caractéristiques des phénomènes naturels, la mise à disposition d'une connaissance complète du phénomène d'intérêt à travers une seule modalité est rarement possible. À titre indicatif, en théorie, un modèle informatisé multimodal consiste en un système susceptible d'intégrer simultanément plusieurs modalités.

Le recours à une approche multimodale permet d'obtenir non seulement des performances accrues mais également davantage de robustesse. Le concept principal qui sous-tend la multimodalité réside dans la *complémentarité*, où chaque modalité fournit au système des informations particulières/supplémentaires qui ne peuvent être obtenues ni déduites auprès d'autres modalités. Du point de vue mathématique, cette information additionnelle est désignée par le terme de diversité [20]. En effet, ce terme consiste à fournir des circonstances favorables au système afin d'en améliorer la robustesse, la performance, la prise de décision ainsi que d'obtenir une vision globale du système.

La classification multimodale a fait l'objet de recherches très actives au cours des dernières années et elle a servi à de nombreuses applications d'intérêt pratique [109], notamment le recours aux techniques d'apprentissage d'ensemble [8, 110]. À titre de rappel, l'apprentissage d'ensemble est une méthode d'apprentissage automatique considérée par la communauté de la reconnaissance des formes comme une tâche complexe [111]. L'apprentissage d'ensemble vise à former plusieurs modèles de base en tant que membres d'un ensemble puis à combiner leurs résultats en une seule sortie afin de parvenir à un modèle prédictif optimal avec des décisions plus précises et plus fiables. La classification d'ensemble se révèle particulièrement utile dans le cas où différents classifieurs sont formés sur diverses parties de l'espace de caractéristiques ou lorsque des ensembles hétérogènes de caractéristiques sont disponibles en vue de les utiliser dans un problème de classification multimodale.

Grâce à la mise à disposition d'informations complémentaires, les données multimodales contribuent habituellement à l'obtention de performances satisfaisantes dans le cadre des tâches de classification. Néanmoins, la combinaison de diverses informations provenant de plusieurs modalités représente un défi majeur, en particulier dans le cas où des données hétérogènes sont concernées. Ainsi, la fusion a pour objectif de mettre en corrélation les éléments de chaque modalité tout en améliorant la qualité de ce qui est exposé et ce, en choisissant de faire apparaître le meilleur de chaque modalité. On distingue deux principaux types d'architecture communément utilisés pour combiner l'information multimodale, à savoir la fusion précoce (fusion au niveau de caractéristiques) et la fusion tardive (fusion au niveau de décisions).

2.2.1 La fusion précoce

La stratégie de fusion précoce consiste à exploiter directement les caractéristiques extraites des modalités à combiner pour créer une représentation commune des caractéristiques d'entrée en vue de prendre la décision finale du système multimodal. Dans ce cas, il n'y a pas de phase de décision intermédiaire prévue pour chaque modalité. Par exemple, dans le cadre de cette étude, cela signifierait qu'aucune décision n'est prise sur la classification de la maladie du glaucome une fois que les caractéristiques sont extraites des images multimodales et fusionnées. La figure 3.5 ci-dessous illustre ce processus.

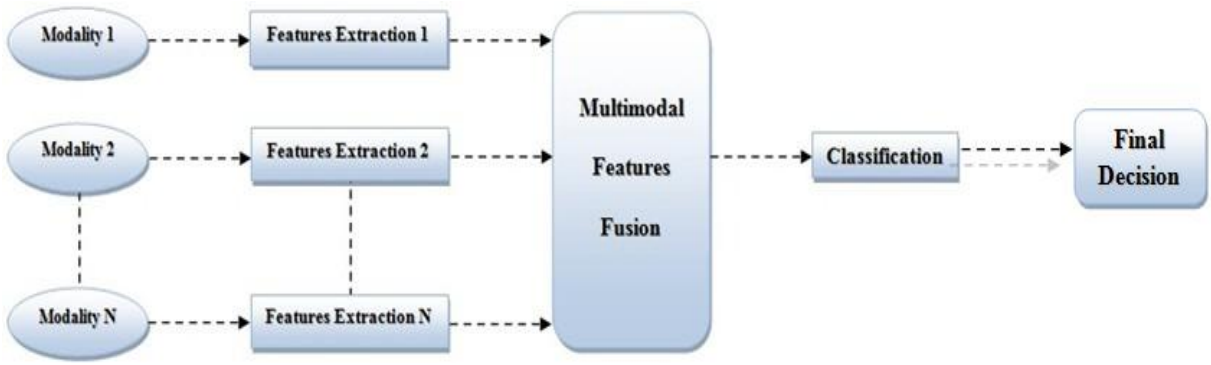

Figure 3.5: Processus général de la stratégie de fusion précoce: la fusion est appliquée directement aux caractéristiques extraites de différentes modalités [8]

2.2.2 La fusion tardive

La stratégie de fusion tardive consiste à fusionner les décisions prises à propos de différentes modalités, plutôt que les caractéristiques extraites directement de ces modalités. Cela permet d'apprendre directement les concepts sémantiques au niveau unimodal. Ainsi, à titre

d'exemple, dans le contexte de cette étude, nous commençons par appliquer des classifieurs séparés pour chaque modalité en envisageant que les modalités sont indépendantes, puis nous fusionnons leurs sorties par le biais d'une méthode de vote, le schéma suivant (Figure 3.6) montre clairement ce principe.

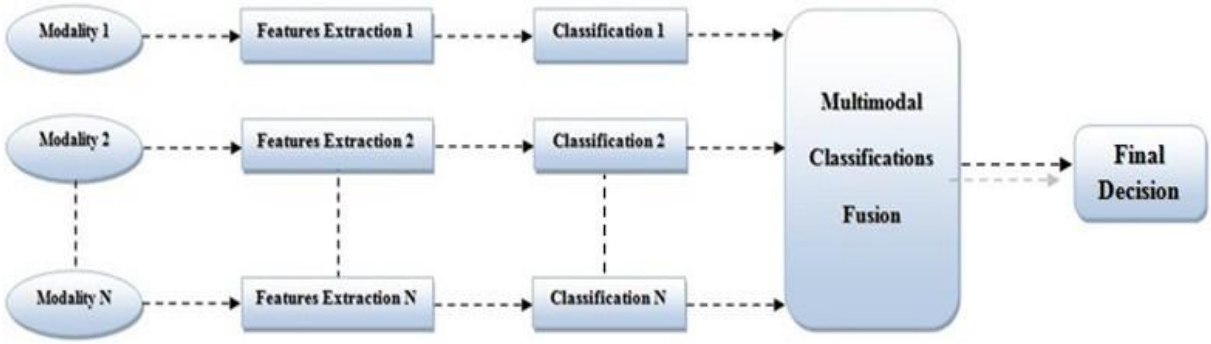

Figure 3.6: Illustration à propos du processus général de la stratégie de fusion tardive: la fusion est appliquée à un ensemble de décisions prises au niveau unimodal [8]

2.2.3 Nouvelle approche de fusion: *fusion hybride*

Dans la présente étude, nous proposons la mise en œuvre d'une nouvelle stratégie de fusion, désignée sous le nom d'approche de *fusion hybride*, dans le but de tirer profit de ces deux techniques de fusion précoce et tardive. D'une part, les différentes modalités sont combinées avant l'apprentissage et, d'autre part, des classifieurs distincts sont utilisés séparément pour chaque combinaison de modalités de manière à assurer concrètement le concept de *multimodalité*. Par la suite, les différents résultats issus de ces classifieurs sont combinés au moyen d'une technique de vote en vue de parvenir à une décision finale de notre système multimodal. Un tel concept se reflète dans la figure 3.7.

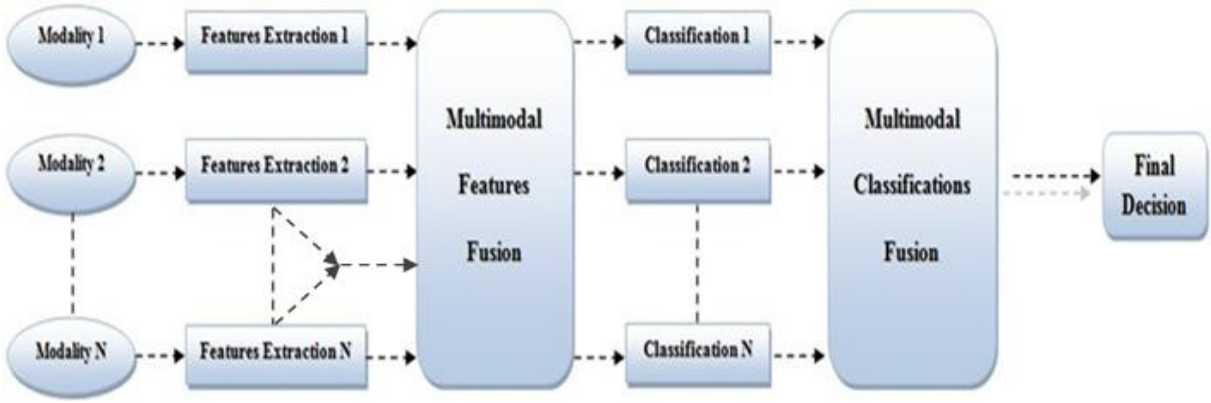

Figure 3.7: Illustration globale du principe de l'approche de fusion hybride proposée [8]

2.3 La méthode proposée

La mise en œuvre de notre réseau par le biais de l'approche multimodale proposée, schématiquement illustrée dans la figure 3.8, est le fruit de plusieurs expérimentations ainsi que d'une étude approfondie de la littérature concernant d'autres tâches de reconnaissance de formes dans le cadre de fusion d'informations.

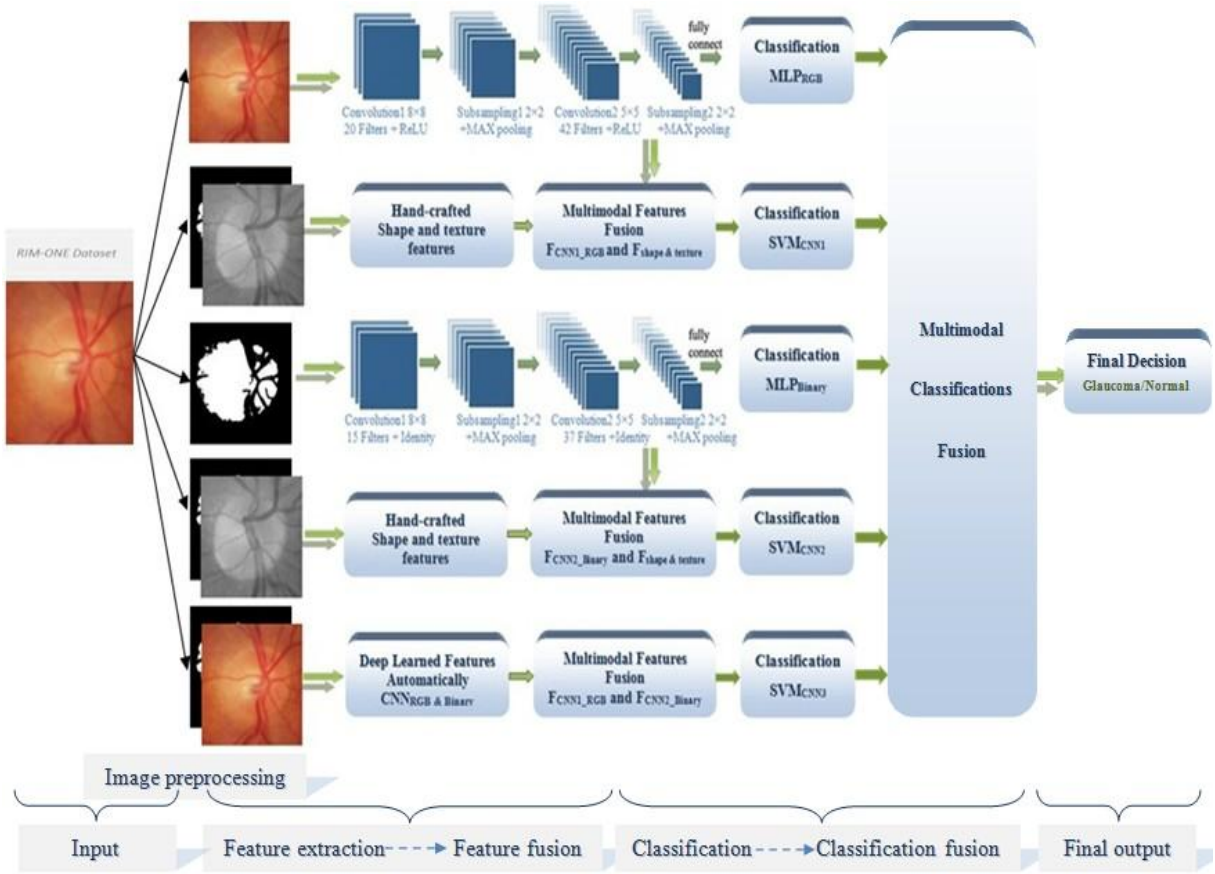

Figure 3.8: Architecture de réseau proposée pour le diagnostic du glaucome [8]

Le système de classification multimodale proposé repose essentiellement sur deux aspects de la multimodalité à savoir, celui du type de ressource ainsi que celui du type caractéristique, en vue de coopérer étroitement aux fins d'obtenir un système très performant.

En premier lieu, nous entraînons deux réseaux *CNNs* sur la base des données multimodales, c'est-à-dire sur la modalité des images couleur RVB et sur celle des images binaires provenant des images du fond de la rétine. En second lieu, nous appliquons le principe de la fusion précoce, consistant à extraire les caractéristiques pertinentes à partir des diverses modalités puis à les fusionner (*bag of features*), de manière à constituer uniquement un vecteur unique sur lequel l'apprentissage se réalise; d'une part, les caractéristiques générées de façon

automatique par le réseau *CNN* sur la base des deux modalités (images *RVB* et images binaires) et celles produites manuellement au moyen de techniques traditionnelles telles que les primitives de forme et de texture (la matrice de cooccurrence de niveau de gris « *GLCM* », les moments Hu et les moments centraux), sont combinées en une représentation multimodale unique en vue de parvenir à une meilleure représentativité possible des caractéristiques. Tandis que, d'autre part, les caractéristiques produites directement à partir des deux réseaux *CNNs* représentant chacune une modalité sont fusionnées dans un seul vecteur multimodal pour une meilleure classification des images.

En dernier lieu, une fois obtenue la représentation multimodale de ces caractéristiques unimodales, il convient de procéder à l'application des classifieurs distincts pour chaque vecteur de caractéristiques multimodales et ce dans le but de fusionner les cinq sorties multimodales de ces classifieurs par le biais des méthodes de vote communément utilisées (*VM*, *VMP* et *VPMP*) de manière à garantir le principe de fusion tardive, et ainsi aboutir au diagnostic final. En outre, le recours aux cinq modèles différents permet d'éviter tout problème d'incompatibilité au moment de la décision finale du système proposé. En ce qui concerne la phase de classification, le classifieur *SVM* est adopté en complément du classifieur *CNN* car celui-ci est réputé être le meilleur séparateur binaire dans le domaine médical.

2.3.1 Prétraitement d'images

La phase préliminaire de prétraitement consiste à obtenir la modalité d'image binaire qui constitue un autre point de vue intéressant de la base de données et ce, en utilisant la méthode *Otsu* [112] (voir Figure 3.9) largement adoptée en raison de ses meilleures performances dans plusieurs domaines liés au traitement de l'image; cette méthode révèle notamment un meilleur contraste entre les structures rétiniennes.

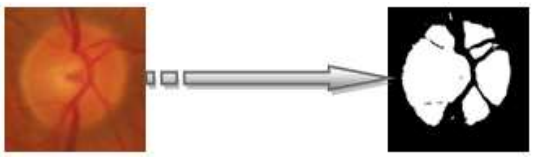

Figure 3.9: Exemple d'une image binaire issue de la base *RIM-ONE* via la technique *Otsu*

En effet, la binarisation représente une étape importante de tout processus de traitement et d'analyse d'images. Cette opération permet de produire deux classes de pixels, généralement représentés par des pixels noirs et des pixels blancs. La binarisation d'une image peut se faire

à l'aide d'un seuil: les pixels dont le niveau de gris est inférieur au seuil deviennent noirs, et ceux qui se trouvent au-dessus de ce seuil prennent une couleur blanche.

En comparant l'image binarisée de bonne qualité avec celle de la source, cela permet d'obtenir une reconnaissance de formes plus précise, étant donné que l'image binarisée ne contient pas de bruit [8, 113]. Ainsi, la méthode d'Otsu figure parmi les techniques de calcul automatique des seuils relatifs à l'inévitable binarisation. Son principe consiste à déterminer le seuil permettant de minimiser la variance pondérée intra-classe tout en maximisant la variance inter-classe [112]. Autrement dit, la méthode d'Otsu essaie de trouver le seuil t séparant de manière optimale l'histogramme en deux segments.

2.3.2 Conception et paramétrage du réseau CNN

L'architecture proposée des deux réseaux CNNs qui se forment sur la base de deux modalités, soit sur la modalité des images couleur RVB et celle des images binaires se compose des six couches dont la structure se présente de la manière suivante: une couche d'entrée E, une couche de convolution C_1, une couche de sous-échantillonnage (*max-pooling*) P_1, une couche de convolution C_2, une couche de sous-échantillonnage (*max-pooling*) P_2, une couche dense FC ainsi qu'une couche de sortie S.

En effet, cette architecture a permis d'éviter la phase d'extraction artisanale des caractéristiques en procédant à l'extraction et à la classification des images simultanément au sein du même réseau, permettant ainsi un diagnostic automatique. Étant donné que les images sont souvent trop volumineuses pour être utilisées directement dans un réseau CNN, toutes les images de la base de données sont donc redimensionnées à une taille de 100×100 de manière à réduire la complexité du calcul tout en garantissant une échelle standard pour toutes les images employées lors de l'apprentissage.

Le modèle proposé est établi au terme de plusieurs tests de performance. Tout commence par la création des blocs de convolution et, après chaque couche de convolution, une opération de normalisation par lots est appliquée pour réduire le nombre de cartes de caractéristiques. L'algorithme de descente de gradient stochastique est employé avec un momentum égal à $0{,}9$. La méthode de régularisation $L2$ est également utilisée avec un seuil égal à $0{,}0005$ en ce qui concerne le poids et les biais.

En outre, un faible taux d'apprentissage est fixé à $0{,}0001$ pour former notre réseau convolutionnel CNN. Les fonctions d'activation *ReLU* et *Identité* sont prises en compte

relativement à la configuration interne des couches de convolution et de mise en commun avec un pas de 1×1 et une taille de noyau équivalente à 2×2. Dans le cas de la couche *FC*, les fonctions *ReLU* et *Cube* sont utilisées, ainsi que la fonction d'erreur quadratique moyenne (*MSE*) pour optimiser la fonction de perte. En dernier point, la fonction *Softmax*, largement utilisée, est adoptée lors du processus de classification. La figure 3.10 schématise le processus d'apprentissage de notre réseau CNN.

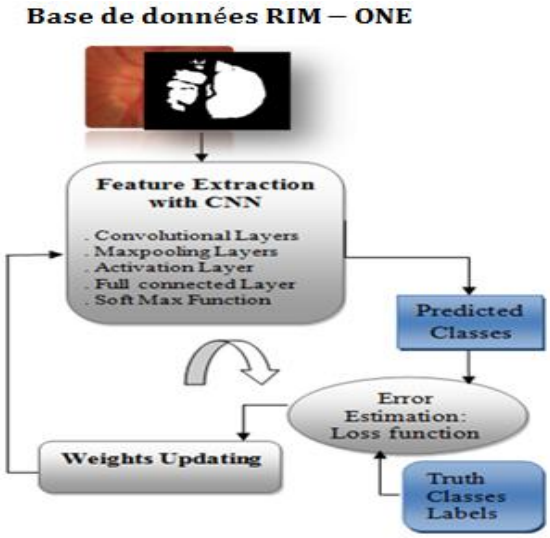

Figure 3.10: Processus d'apprentissage supervisé du classifieur CNN

2.3.3 Complexité de calcul

Depuis peu, il y a eu un intérêt croissant pour accélérer l'exécution des réseaux CNNs. le temps d'exécution réel peut être sensible aux mises en œuvre et aux équipements. Les modèles proposés dans le cadre de cette étude présentent un coût de calcul abordable et peu coûteux, qui ne prend qu'une journée de formation au moyen d'un seul processeur. Toutefois, une fois que nos modèles sont préformés (hors ligne), la notion de coût de calcul n'intervient pas durant la phase de classification en temps réel (en ligne). La figure 3.11 suivante indique qu'à partir d'un certain nombre d'itérations, le modèle proposé se converge ; cette figure représente le score du modèle en fonction de l'itération des deux *CNNs*, soit la valeur de la fonction de perte sur le mini-lot actuel.

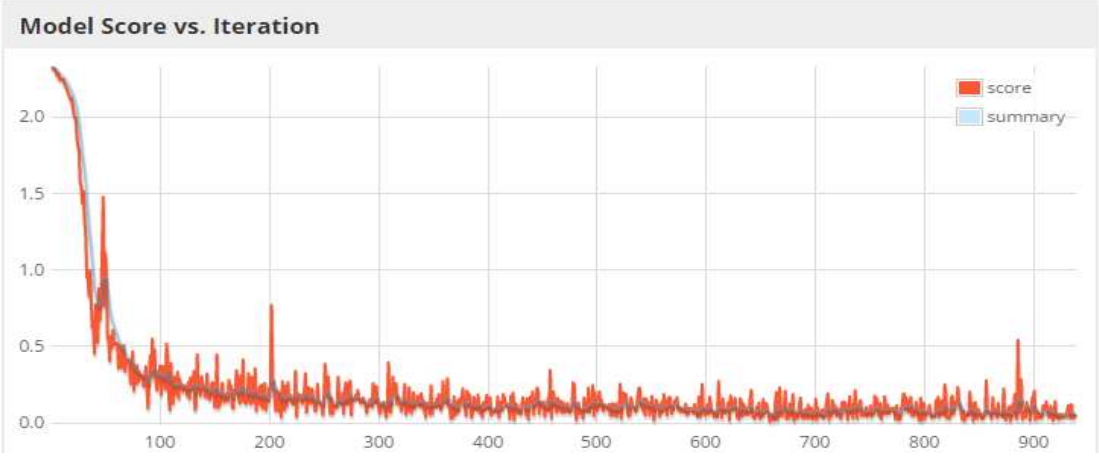

Figure 3.11: Score du modèle versus itération

Les principaux paramètres du modèle et les informations d'apprentissage de l'architecture *CNN* proposée sont décrits dans les tableaux ci-dessous.

Tableau 3.7: Informations d'apprentissage du modèle $CNN1_{RVB}$

Model Type	MultiLayerNetwork
Layers	6
Total Parameters	431080
Last Update	2018-05-12 23:57:39
Total Parameter Updates	938
Updates/sec	3,47
Examples/sec	111,11

Tableau 3.8: Informations d'apprentissage du modèle $CNN2_{Binaire}$

Model Type	MultiLayerNetwork
Layers	6
Total Parameters	431080
Last Update	2018-05-12 23:39:38
Total Parameter Updates	938
Updates/sec	3,36
Examples/sec	107,38

La limitation de la complexité du réseau est un moyen de comprendre l'impact de certains facteurs tels que la profondeur (nombre de couches), la largeur (nombre de filtres), la taille des filtres ainsi que les progrès des architectures dans la conception du réseau. Ces différents facteurs permettent d'évaluer la fiabilité du réseau *CNN* tandis que leur accroissement

engendre une augmentation des coûts de calcul. Le coût de calcul de la couche de convolution correspond à $O(F_I MNmnF_O)$, où $M \times N$ représente la taille de la carte des caractéristiques pour chaque entrée, $m \times n$ désigne la taille du noyau de convolution et F_I, F_O sont respectivement les canaux d'entrée et de sortie dans une couche. Le coût de calcul de la couche de *max-pooling* est souvent très faible au regard de celui de la couche de convolution, qui se définit comme suit: $O(F_I MN)$ [114]. Pour finir, la méthode proposée offre à la fois une reconnaissance précise, dont la complexité de calcul reste minime, et un coût temporel acceptable.

2.3.4 Extraction de caractéristiques au moyen de techniques traditionnelles

Dans le domaine de la reconnaissance des formes, les caractéristiques constituent les propriétés mesurables d'un phénomène physique observé. L'extraction des caractéristiques discriminantes représente une étape fondamentale lors du processus de reconnaissance préalable à la classification. Cette phase d'extraction des caractéristiques consiste à déterminer un ensemble de mesures afin de représenter chaque classe de manière aussi unique que possible tout en réduisant la notion de dimensionnalité. Ainsi, la performance des systèmes de reconnaissance dépend en grande partie du choix des descripteurs utilisés ainsi que des techniques associées à leur extraction.

De nombreux descripteurs sont employés dans le cadre des systèmes de reconnaissance d'images par le contenu et plus particulièrement dans les systèmes d'aide au diagnostic en vue de décrire des formes telles que des descripteurs de couleurs, de textures et de formes. La majorité des travaux proposés dans la littérature ont recours au concept de combinaison de divers descripteurs des caractéristiques [115-117]. Cette combinaison a montré que les résultats obtenus permettent de fournir une meilleure précision de diagnostic.

Dans le cadre de cette étude, concernant les méthodes traditionnelles d'extraction de caractéristiques, nous nous appuyons sur deux familles de caractéristiques hétérogènes couramment utilisées, à savoir la texture et la forme, dans le but de les concaténer conjointement à celles générées automatiquement par le réseau CNN afin d'assurer une meilleure représentation vectorielle des caractéristiques de l'image médicale. En termes de caractéristiques de texture, la matrice de cooccurrence de niveau de gris « *GLCM* » est employée, tandis que les deux méthodes d'extraction des caractéristiques de forme, *Moments centraux* et *Moments Hu*, sont appliquées [8, 80, 118, 119].

2.3.4.1 Matrice de cooccurrence de niveau de gris « GLCM »

Du fait de leur richesse en informations relatives aux textures, les matrices de co-occurrence constituent la méthode la plus connue et la plus utilisée permettant d'extraire ces caractéristiques de texture [120, 121]. Les matrices de cooccurrences consistent à mesurer la probabilité d'apparition de paires de valeurs de pixels situées à une certaine distance dans l'image. Il est basé sur le calcul de la probabilité $p\ (i, j, d, \theta)$ qui représente le nombre de fois qu'un pixel de niveau de couleur i apparaît à une distance relative d d'un pixel de niveau de couleur j et selon une orientation θ donnée. Les directions angulaires θ conventionnellement utilisées sont 0, 45, 90 et 135 degrés. Cette matrice caractérise les motifs identifiables en niveaux de gris d'une région de pixels. Comme les matrices de cooccurrences comptent une très grande quantité d'informations difficiles à exploiter directement pour caractériser les textures. Haralick [122] a proposé les quatorze premiers paramètres, caractérisant les textures, issues de ces matrices. Récemment, dans le diagnostic médical, seules les treize premières caractéristiques les plus appropriées sont couramment utilisées [123, 124] et qui sont prises en compte dans notre étude [8].

2.3.4.2 Moments centraux

Les moments sont des quantités scalaires utilisées pour décrire une fonction et capturer ses caractéristiques importantes. La notion de moments en mathématiques, sont des projections d'une fonction sur une base polynomiale; différents systèmes de moments peuvent être reconnus selon la base polynomiale utilisée. Les moments centraux font désormais partie des descripteurs de formes les plus utilisés dans de nombreux domaines [125, 126] qui ont fait preuve de performances supérieures. Un moment central est un moment d'une distribution de probabilité d'une variable aléatoire sur la moyenne de la variable aléatoire [8].

2.3.4.3 Moments Hu

Une reformulation des moments précédemment décrits s'avère nécessaire pour permettre l'invariance de la rotation, la rotation Hu-moment invariants (*HMI*) a été proposée par Hu [127]. Hu [127] a obtenu ces expressions grâce à des invariants algébriques effectués à la fonction génératrice du moment sous une transformation de rotation. Un ensemble

d'expressions de moments centralisés non linéaires constitue les *IHMs*, qui sont complètement orthogonales (c'est-à-dire de rotation) invariantes [8].

En vue de calculer les moments centraux et les moments Hu permettant d'extraire les caractéristiques de forme, il convient de convertir les images couleur en images binaires - dans ce but, la méthode *Otsu* a été adoptée, telle que présentée auparavant. La matrice de cooccurrence est calculée à partir d'images en niveaux de gris à l'aide de l'outil *ImageJ*[4].

2.3.5 Fusion de classification multimodale

En effet, en ce qui concerne la reconnaissance des formes, il n'existe pas de modèle unique pour tous les problèmes ni de méthode unique applicable à tous ces derniers ; autrement dit, il n'existe pas de "meilleur" classifieur capable de traiter/apprendre une distribution quelconque de données d'apprentissage. A ce titre, il n'a pas été possible de mettre en évidence l'indéniable supériorité d'une méthode de classification sur une autre et d'un module d'extraction de caractéristiques par rapport à un autre, d'où notre intérêt pour l'apprentissage d'ensemble.

Comme déjà mentionné dans la section 2.2, nous faisons appel à l'approche de fusion hybride proposée, c'est-à-dire la concaténation des caractéristiques de différentes modalités (*fusion précoce*) ainsi que la classification multimodale (*fusion tardive*). Dans le but de tirer avantage de la complémentarité pouvant exister entre les classifieurs, d'une part, et de l'ensemble des caractéristiques multimodales générées conformément aux méthodes adoptées (*CNN*, *GLCM*, *Moments centraux* et *Moments Hu*), d'autre part, cinq modèles sont mis en œuvre, soit MLP_{RVB}, SVM_{CNN1}, $MLP_{Binaire}$, SVM_{CNN2} et SVM_{CNN3} (voir Figure 3.8).

La présente étude adopte la méthode d'ensemble parallèle dans laquelle les cinq apprenants de base sont générés séparément et de manière parallèle. La principale motivation qui nous pousse à utiliser une telle méthode est d'exploiter l'indépendance entre les cinq classifieurs de base, dans la mesure où la fusion permet de réduire considérablement l'erreur. Une fois que les algorithmes du pool (les multiples apprenants de base) sont formés, il faut recourir à une technique de combinaison appropriée en vue de combiner leurs sorties (d'apprentissage) sous une forme unique en tant que classifieur final.

[4] http://imagej.net/Welcome

En dépit du grand nombre de méthodes de combinaison figurant dans la littérature, trois d'entre elles seulement sont largement appliquées et présentent un potentiel d'amélioration considérable dans de nombreuses applications d'apprentissage d'ensemble, à savoir le combineur linéaire, le combineur produit et le combineur vote.

Dans le cadre de la mise en œuvre de la décision finale relative au système multimodal suggéré, des techniques avancées de fusion basées sur le vote sont utilisées. De telles stratégies sont déjà développées préalablement (voir Section 1.3) et comprennent le vote à la majorité (*VM*), le vote à la majorité pondérée (*VMP*) et le vote pondéré meilleur-pire (*VPMP*).

2.4 Résultats expérimentaux

Dans le but d'évaluer la méthode de classification multimodale proposée concernant la maladie du glaucome, la technique de $k - bloc$ validation croisée est employée pour $k = 10$. La réalisation de ce travail se fait grâce à la plateforme *Deeplearning4j*[1]. Les sous-sections qui suivent décrivent de manière détaillée les expériences menées et les résultats obtenus.

2.4.1 Description de la base de données utilisée

Dans le cadre de la méthode proposée en matière de diagnostic du glaucome, un ensemble composé de 455 images issues de la base de données ouverte *Retinal Image Database for Optic Nerve Evaluation* (*RIM-ONE*) est utilisé, dont 200 images représentent la maladie du glaucome tandis que 255 autres sont des images de type normal; celles-ci étant centrées sur l'*ONH* et ayant un champ de vision (*FOV*) égal à 34°. La figure 3.12 illustre des exemples d'images relatives au fond d'œil rétinien.

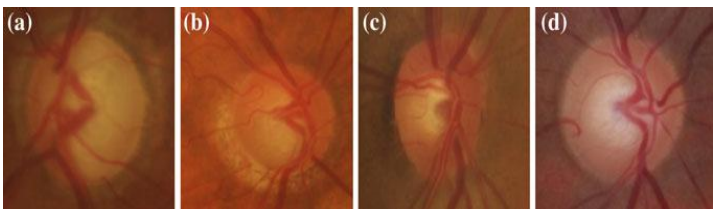

Figure 3.12: Exemples d'images du fond d'œil de la base de données RIM-ONE: $(\boldsymbol{a})$ et $(\boldsymbol{b})$ glaucome, $(\boldsymbol{c})$ et $(\boldsymbol{d})$ normal

2.4.2 Les mesures d'évaluation

La performance du système suggéré repose sur un ensemble de critères d'évaluation communément utilisés dans le domaine médical, et qui sont décrits de la manière suivante:

$$\text{Précision (Acc)} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{3.8}$$

$$\text{Sensibilité (Sen)} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.9}$$

$$\text{Spécificité (Spe)} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{3.10}$$

$$Valeur\ prédictive\ positive\ (VPP) = \frac{VP}{VP + FP} \tag{3.11}$$

$$Valeur\ prédictive\ négative\ (VPN) = \frac{VN}{VN + FN} \tag{3.12}$$

$$Coefficient\ de\ corrélation\ Matthews\ (CCM) = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.13}$$

Où, Vrais Positifs (VP) désignent les échantillons de personnes malades présentant un test positif, Vrais Négatifs (VN) sont des échantillons de personnes non malades ayant un test négatif, Faux Positifs (FP) représentent les échantillons de personnes non malades avec un test positif, et Faux Négatifs (FN) concernent les échantillons de personnes malades dont le test est négatif.

2.4.3 Résultats et discussions

Ce travail se base essentiellement autour de cinq modèles différents, à savoir: le CNN1_{RVB} reposant sur les images couleur du fond d'œil (*images originales de l'ensemble de données RIM-ONE*), le SVM_{CNN1} intégrant un ensemble de caractéristiques de type artisanal (*GLCM, Moments centraux, Moments Hu*) et automatique (*carte de caractéristiques générée via* $CNN1_{RVB}$), le $\text{CNN2}_{\text{Binaire}}$ se basant sur les images binaires (*images originales RIM-ONE converties au moyen de l'algorithme Otsu*), le SVM_{CNN2} utilisant toutes les caractéristiques à la fois (*GLCM, moments centraux, moments Hu et carte des caractéristiques générée en* $CNN2_{Binaire}$) ainsi que le SVM_{CNN3} qui se fonde sur les caractéristiques fournies par les

modèles $CNN1_{RVB}$ et $CNN2_{Binaire}$. Les tableaux suivants récapitulent la performance individuelle obtenue pour chaque classifieur.

Le tableau 3.9 (MLP_{RVB}) illustre que le modèle $CNN1_{RVB}$ a reconnu de manière correcte 197 images de glaucome comme étant des images glaucomateuses, 252 images saines sont correctement classées comme des images non glaucomateuses. En résumé, 449 images sont étiquetées de manière précise, résultant en une précision (Acc) de 98.68% dont la sensibilité (Sen) est de 98.50%, la spécificité (Spe) est de 98.82%, la valeur prédictive positive (VPP) est de 98.50%, la valeur prédictive négative (VPN) est de 98.82% et le coefficient de corrélation de Matthews (CCM) est de 97.32%.

Tableau 3.9: Matrice de confusion des résultats obtenus par le modèle $CNN1_{RVB}$

	Glaucome	Sain
Glaucome	197	3
Sain	3	252

Le tableau 3.10 ($MLP_{Binaire}$) montre que, selon le modèle $CNN2_{Binaire}$, 199 images de glaucome sont correctement détectées comme des images glaucomateuses et 253 images saines sont correctement classées comme des images non glaucomateuses. Pour résumer, 452 images sont correctement étiquetées, ce qui se traduit par une précision (Acc) de 99.34%, une Sen de 99.50%, une Spe de 99.22%, une valeur VPP de 99.00%, une valeur VPN de 99.61% et un coefficient CCM de 98.66%.

Tableau 3.10: Matrice de confusion des résultats fournis d'après le modèle $CNN2_{Binaire}$

	Glaucome	Sain
Glaucome	199	1
Sain	2	253

Dans le tableau 3.11 (SVM_{CNN1}), il est démontré que 198 images de glaucome sont correctement identifiées en tant qu'images glaucomateuses dans le cadre du modèle SVM_{RVB} et que 253 images saine/normales sont proprement classées en tant qu'images non glaucomateuses. De manière générale, 451 images sont correctement étiquetées, ce qui aboutit à une précision (Acc) de 99.12%, une Sen de 98.50%, une Spe de 99.61%, une valeur VPP de 99.49%, une valeur VPN de 98.83% et un coefficient CCM de 98.22%.

Tableau 3.11: Matrice de confusion des résultats produits selon le modèle SVM_{RVB}

	Glaucome	Sain
Glaucome	198	2
Sain	2	253

Comme indiqué dans le tableau 3.12 (SVM_{CNN2}), 198 images de glaucome sont correctement classées dans la catégorie des images glaucomateuses par le modèle $SVM_{Binaire}$, alors que 255 images normales sont proprement classées dans la catégorie des images non glaucomateuses. Au total, 453 images sont correctement étiquetées, ce qui donne une précision (Acc) de 99.56% avec une Sen de 99.00%, une Spe de 100%, une valeur VPP de 100%, une valeur VPN de 99.22% et un coefficient CCM de 99.11%.

Tableau 3.12: Matrice de confusion des résultats obtenus avec le modèle $SVM_{Binaire}$

	Glaucome	Sain
Glaucome	198	2
Sain	0	255

Le tableau 3.13 (SVM_{CNN3}) illustre que 200 images de glaucome sont identifiées de manière correcte comme étant des images glaucomateuses en fonction du modèle $SVM_{RVB\&Binaire}$, tandis que 252 images de type sain/normal sont correctement classées comme étant des images non glaucomateuses. Pour récapituler, 452 images sont correctement étiquetées, résultant en une précision (Acc) de 99.34% avec une Sen de 100%, une Spe de 98.82%, une valeur VPP de 98.52%, une valeur VPN de 100% et un coefficient CCM de 98.67%.

Tableau 3.13: Matrice de confusion des résultats donnés suivant le modèle $SVM_{RVB\&Binaire}$

	Glaucome	Sain
Glaucome	200	0
Sain	3	252

En outre, une étude comparative portant sur les performances du classifieur SVM au moyen de diverses combinaisons possibles de familles de caractéristiques a été menée, dont la figure 3.13 illustre les résultats ainsi obtenus.

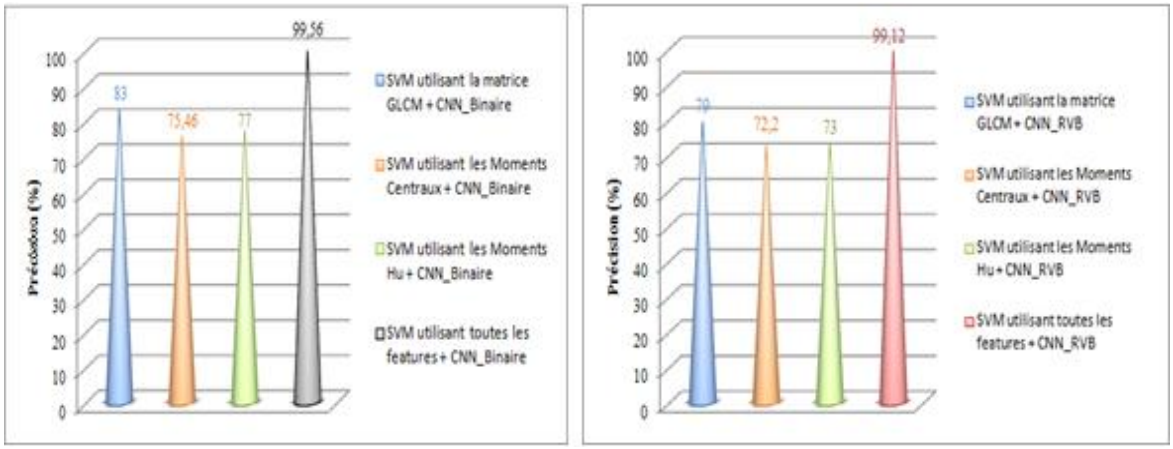

Figure 3.13: Analyse comparative des résultats obtenus au moyen de combinaisons différentes de caractéristiques

À la suite de la comparaison des résultats obtenus, nous constatons que la combinaison des trois techniques d'extraction artisanale des caractéristiques (*GLCM*, *Moments centraux* et *Moments Hu*) aux caractéristiques produites automatiquement au moyen du réseau *CNN* (*pour les deux modalités*) est largement supérieure/préférable au recours à la combinaison séparée de ces trois dernières techniques avec les caractéristiques fournies par *CNN*, ce qui permet d'améliorer davantage la performance du diagnostic.

A cet effet, les principaux paramètres à régler concernant le classifieur SVM sont définis dans les termes suivants: Noyau de fonction de base radiale (*RBF*) avec une valeur $Gamma$ de 0.5 et une valeur du paramètre C égale à 4. L'approche proposée est également évaluée sur la base des courbes *ROC* [128]. Les courbes ROC de nos modèles sont tracées sur la figure 3.14.

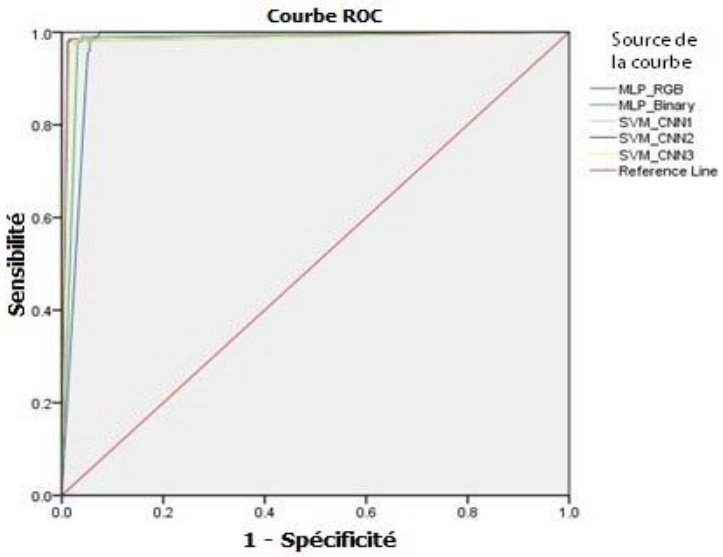

Figure 3.14: Courbes ROC correspondant aux 5 modèles utilisés [8]

En observant que tous les points de courbe des cinq classifieurs mis en œuvre sont positionnés dans la moitié supérieure de l'espace ROC, résultant en une bonne courbe ROC, en particulier pour les modèles basés sur le classifieur SVM. Le tableau 3.14 ci-après récapitule à la fois les résultats obtenus ainsi que les valeurs de l'aire sous la courbe «*AUROC*» correspondant à chaque modèle mis en œuvre.

Tableau 3.14: Synthèse au sujet des performances obtenues selon les cinq modèles utilisés

Modèle	Acc(%)	Sen(%)	Spe(%)	VPP(%)	VPN(%)	CCM(%)	AUROC(%)
MLP_{RVB}	98.68	98.50	98.82	98.50	98.82	97.32	97.30
$MLP_{Binaire}$	99.34	99.50	99.22	99.00	99.61	98.66	97.90
SVM_{CNN1}	99.12	98.50	99.61	99.49	98.83	98.22	98.20
SVM_{CNN2}	99.56	99.00	100	100	99.22	99.11	98.70
SVM_{CNN3}	99.34	100	98.82	98.52	100	98.67	98.40

Au cours de cette étude, plusieurs expérimentations sont effectuées concernant les performances du réseau CNN, en particulier en matière de nombre d'époques et de la fonction d'activation adoptée. La figure 3.15 met en évidence l'impact du nombre d'époques sur la précision au niveau de deux classifieurs $CNN1_{RVB}$ et $CNN2_{Binaire}$ conformément à la fonction d'activation utilisée.

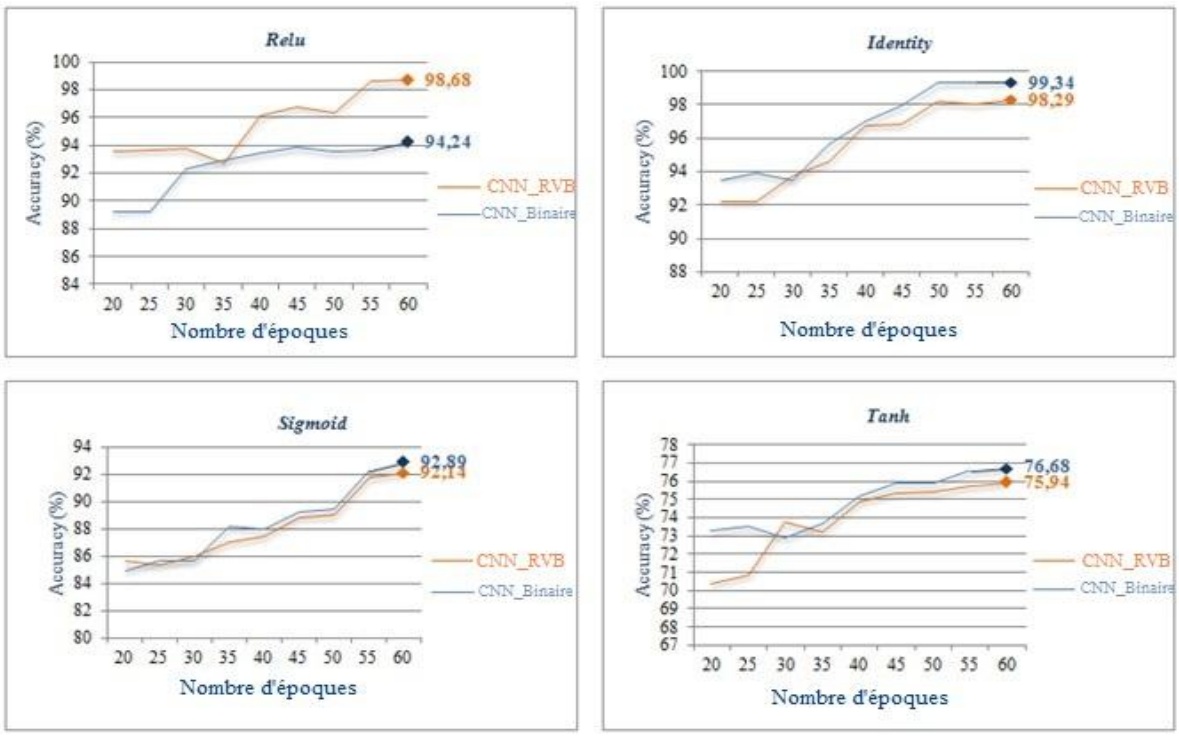

Figure 3.15: Représentation graphique de la précision (Acc) des deux modèles $CNN1_{RVB}$ et $CNN2_{Binaire}$ par rapport aux différentes époques selon les fonctions d'activation utilisées [8]

Sur un plan général, il est apparu clairement que l'augmentation dans le nombre d'époques se traduisait par une amélioration du taux de précision pour les deux réseaux CNN ($CNN1_{RVB}$ et $CNN2_{Binaire}$) quelle que soit la fonction d'activation « *ReLU*, *Identité*, *Sigmoid* ou *Tanh* » considérée dans ce travail.

Conformément à ce graphique, on peut conclure que les classifieurs $CNN1_{RVB}$ et $CNN2_{Binaire}$ atteignent leur performance maximale lorsque le nombre d'époques équivaut à soixante dans le cas des quatre fonctions d'activation utilisées, alors que la précision est pratiquement stable au bout de cinquante-cinq époques. Par ailleurs, on constate également que le modèle $CNN1_{RVB}$ présente un taux de précision supérieur avec la fonction *ReLU* (98.68%), tandis que le modèle $CNN2_{Binaire}$, grâce à la fonction *Identité*, surpasse toutes les autres fonctions d'activation tout en montrant une précision de 99.38 %, ce qui permet d'obtenir des résultats plus satisfaisants.

Aux fins de renforcer la sensibilité et la spécificité de ce système multimodal ainsi que la performance globale du diagnostic, d'une part, et afin de parvenir à la décision finale de l'approche suggérée, d'autre part, trois stratégies d'agrégation sont appliquées de manière à ce que les résultats soient plus précis et plus fiables: le vote à la majorité (*VM*), le vote à la majorité pondérée (*VMP*) et le vote pondéré meilleur-pire (*VPMP*). Le tableau 3.15 reprend les résultats finaux ainsi obtenus et la figure 3.16 qui suit illustre les courbes ROC.

Tableau 3.15: Les résultats fournis par les trois méthodes de fusion utilisées

Modèle	Acc(%)	Sen(%)	Spe(%)	VPP(%)	VPN(%)	CCM(%)	AUROC(%)
VM	99.12	99.00	99.22	99.00	99.22	98.22	98.60
VMP	99.56	99.50	99.61	99.50	99.61	99.11	98.80
VPMP	99.78	99.50	100	100	99.61	99.55	99.20

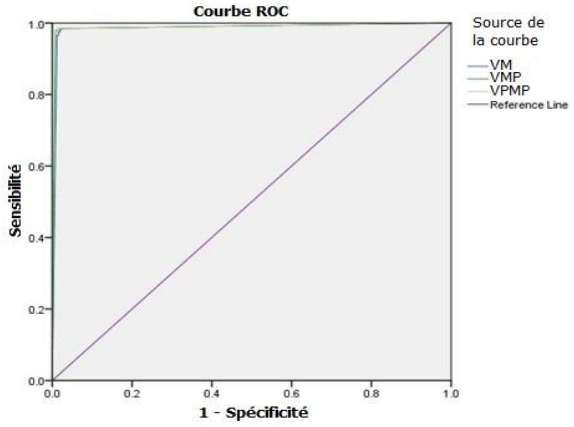

Figure 3.16: Courbes ROC pour les trois approches d'agrégation adoptées [8]

Sur la base des résultats exposés ci-dessus, nous jugeons que la règle du vote pondéré meilleur-pire (*VPMP*) est celle qui offre les meilleurs résultats, et ce avec un niveau de précision particulièrement satisfaisant, en comparant avec les résultats de pointe utilisant la même base de données (*RIM-ONE*).

Au terme de l'analyse des résultats expérimentaux, nous sommes parvenus à la conclusion que la combinaison de classifieurs par le biais de la méthode VPMP est plus préférable sur le plan de l'apprentissage séparé de ces classifieurs. De même, la combinaison de plusieurs caractéristiques multimodales a prouvé à quel point elle est particulièrement efficace en matière de diagnostic du glaucome.

Le grand atout du présent travail réside dans le fait qu'il permet de concevoir un système d'aide à la décision pour le glaucome dont les performances surpassent celles des autres modèles disponibles dans la littérature, et cela grâce à une architecture multimodale basée sur l'approche de fusion hybride proposée. Ainsi, le tableau 3.16 met en évidence les principales caractéristiques indiquées dans la littérature en comparant les performances obtenues au moyen de la méthode suggérée avec celles des plus importantes approches existant dans la littérature en matière de diagnostic du glaucome.

Tableau 3.16: Étude comparative de la performance du système proposé avec d'autres travaux relatifs au diagnostic du glaucome triés par ordre croissant en termes de précision (Acc)

Référence des travaux	Technique d'extraction	Méthode de Classification	Nombre d'images	Acc(%)	Sen(%)	Spe(%)
Chen et al. [97]	Six couches CNN	-	1,676	-	-	-
Orlando et al. [100]	Huit couches CNN	-	101	-	-	-
Bock et al. [88]	Intensités des pixels bruts, FFT, B-spline et ACP	SVM	575	80.00	73.00	85.00
Chai et al. [98]	CNN à deux branches (cinq couches Conv.)	-	3,554	81.69	-	-
Nayak et al. [129]	Morphologique et Seuillage	RNA	61	90.00	100	80.00
Deepti et al. [131]	GLCM	ART/RNA	100	90.00	-	-
Acharya et al. [130]	HOS, GLCM et GLRLM	Forêts Aléatoires (FA)	60	91.70	-	-
Kevin et al. [132]	HOS	NB/SVM	272	92.60	100	92.00

Noronha et al. [83]	HOS cumulant et ADL	Naïve Bayes (NB)	272	92.65	100	92.00
Dua et al. [85]	Ondelettes DWT	SMO	60	93.00	-	-
Rajendra et al. [133]	Filtre de Gabor	SVM	510	93.10	89.73	96.20
Acharya et al. [138]	Transformée de Gabor et entropie	SVM	510	93.10	89.75	96.20
Kolar & Jan [82]	Dimension fractale (DF)	SVM	30	93.80	-	-
Nagarajan et al. [134]	MVEP	RNA	399	94.00	95.00	94.00
Ashish et al. [135]	CDR, NRR et des quadrants ISNT	SVM & RNA	67	94.00	100	-
Zilly et al. [99]	CNN, transformée de Hough et entropie	-	155	94.10	92.30	95.60
Issac et al. [137]	Morphologique (CDR, NRR, BV et IQ)	SVM & RNA	67	94.11	100	90.00
Singh et al. [91]	Ondelettes DWT	FA et RNA / SVM et k-ppv	63	94.70	-	-
Mookiah et al. [84]	HOS et Ondelettes DWT	SVM	60	95.00	93.33	96.67
Maheshwari et al. [92]	VMD, entropie et DF	LS-SVM	488	95.19	93.62	96.71
Raghavendra et al. [89]	RT, MCT et GIST	SVM	1000	97.00	97.80	95.80
Kausu et al. [93]	Morphologique et Ondelettes DWT	MLP	86	97.67	98.00	97.10
Simonthomas et al. [136]	Haralick (GLCM)	K-ppv	60	98.00	-	-
Raghavendra et al. [101]	Dix-huit couches CNN	ADL	1,426	98.13	98.00	98.30
Méthode proposée [8]	Fusion multimodale des caractéristiques: GLCM, Moments centraux, Moments Hu et CNN profond	Fusion multimodales des classifications (CNN_{MLP}+SVM) en utilisant la stratégie VPMP	455	99.78	99.50	100

2.5 Conclusion

Dans le contexte de la présente étude, nous avons introduit le concept de *multimodalité* en présentant une nouvelle méthode de classification multimodale permettant le diagnostic précoce de la maladie du glaucome sur la base à la fois de données multimodales ainsi que de caractéristiques multimodales issues d'images du fond rétinien (*RIM-ONE*).

Cette approche repose également en grande partie sur le concept de la fusion hybride proposée dans le cadre de cette contribution, tout en recourant aux réseaux *CNN*s et au classifieur *SVM*. L'un des principaux avantages de l'utilisation de *CNN* est sa capacité potentielle à fournir de manière automatique les caractéristiques et les informations les plus pertinentes/utiles.

La combinaison de ces dernières ainsi que des caractéristiques traditionnelles (*GLCM*, *Moments centraux* et *Moments Hu*) a permis de mettre en place un système extrêmement robuste et précis. L'orientation future visée serait la possibilité d'améliorer davantage notre système en y intégrant également d'autres types de modalités, dont principalement des images de carte topographique, soit une nouvelle interprétation de la rétine, ainsi que des informations textuelles relatives au patient telles que l'âge, la myopie, avec ou sans diabète, etc.

Par ailleurs, les résultats finaux obtenus au cours de la phase expérimentale de cette étude démontrent clairement que le recours à la combinaison de classifieurs par le biais de la règle d'agrégation *VPMP* constitue une solution plus avantageuse en termes de performances (Acc de 99.78%, Sen de 99.50% et Spe de 100%) par rapport aux techniques *VM* et *VMP*, surpassant ainsi toutes les autres principales études menées dans ce domaine qui nous permettent de poursuivre dans cette voie de recherche.

Depuis peu, le thème relatif à l'apprentissage déséquilibré devient incontournable et de plus en plus présent dans le domaine de l'apprentissage automatique en relation avec le diagnostic médical, et ce en raison du fait que les bases médicales soient généralement non équilibrées en termes de données d'apprentissage, ce qui peut générer des résultats de classification erronés. Ce défi sera abordé dans le chapitre suivant par le biais de la mise en œuvre d'une nouvelle approche d'ensemble multimodale « *méta-classification* » portant sur une technique de ré-échantillonnage optimisée, ainsi que par l'analyse des différents paradigmes d'agrégation de pointe.

CHAPITRE 04

APPRENTISSAGE MULTIMODAL DÉSÉQUILIBRE DANS LE CADRE DU DIAGNOSTIC PRÉCOCE DU DIABETE PAR LE BIAIS D'UNE TECHNIQUE DE RÉ-ÉCHANTILLONNAGE AMÉLIORÉE

Chapitre 04. Apprentissage Multimodal Déséquilibré dans le cadre du Diagnostic Précoce du Diabète par le biais d'une Technique de Ré-échantillonnage Améliorée

1 Introduction

L'apprentissage déséquilibré - ou l'apprentissage à partir de données déséquilibrées - représente un problème courant en rapport avec les bases de données médicales qui sont souvent déséquilibrées de sorte qu'il est également considéré comme un autre challenge en matière de l'apprentissage automatique, en particulier en ce qui concerne l'apprentissage supervisé. Récemment, la notion de déséquilibre de classe a fait son apparition au sein de divers domaines d'application présentant un intérêt pratique, y compris le dépistage des maladies [139], la prédiction d'événements rares [140] et le filtrage du spam [141], où il existe plus d'échantillons disponibles pour certaines catégories que pour d'autres.

Par exemple, dans le cas d'un problème de classification binaire, la classe minoritaire dispose de données d'apprentissage beaucoup moins représentées par rapport à celles de la classe majoritaire, avec pour conséquence une perturbation au niveau des algorithmes d'apprentissage automatique [142, 143]. Le schéma ci-dessous (Figure 4.1) illustre un exemple au sujet des données déséquilibrées.

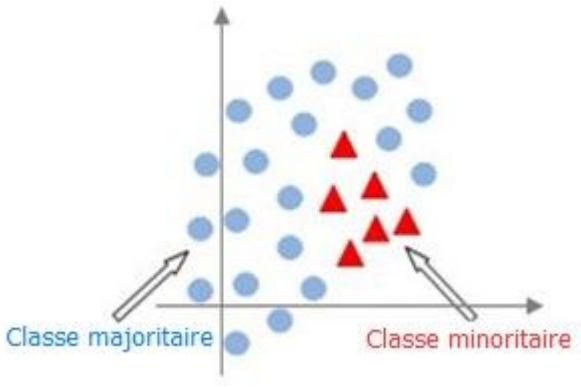

Figure 4.1: Exemple illustrant le concept relatif aux données déséquilibrées

En fait, la plupart des méthodes d'apprentissage automatique tendent toujours à négliger la classe minoritaire, alors que dans la majorité des situations, cette classe représente le point le plus intéressant. Cela est dû au fait que la frontière de séparation ou la limite de décision converge au profit de la classe majoritaire; ce qui signifie que les algorithmes d'apprentissage automatique apprennent uniquement à prédire les classes récurrentes de manière à ce que les classifieurs peuvent aboutir ainsi à des résultats de classification erronés, spécialement en ce qui concerne l'ensemble de validation.

À cette fin, la mise en œuvre d'un module d'équilibrage des ensembles de données serait intéressante de manière à éviter les erreurs d'apprentissage ainsi qu'une classification biaisée. Deux aspects sont abordés au moyen de ce module, soit le traitement au *niveau des données* ainsi qu'au *niveau des algorithmes*. Néanmoins, la majorité des approches traitent ce problème principalement au niveau des données en termes de modification/équilibrage de l'ensemble de données d'apprentissage à travers la mise en pratique des techniques de ré-échantillonnage (sur- et/ou sous-échantillonnage), telles que les stratégies *RESAMPLE* et *SMOTE (Synthetic Minority Over-sampling Technique)*, qui figurent parmi les méthodes largement adoptées permettant de contourner tout problème relatif aux données déséquilibrées [139, 140, 142, 143].

À titre d'exemple relatif au diagnostic du diabète, Nnamoko et Korkontzelos [139] ont mis au point une approche de prétraitement de données à deux niveaux; d'une part, par le traitement des valeurs aberrantes au moyen de l'algorithme de l'intervalle interquartile (*IQR*). D'autre part, la technique *SMOTE* est adoptée pour faire face au problème de déséquilibre des données. Dans le cadre de cette étude, les auteurs ont analysé plusieurs classifieurs, à savoir SVM-RBF, NB, C4.5 et RIPPER, en utilisant la méthode de validation croisée à $10\ blocs$. Les résultats expérimentaux ont indiqué que l'algorithme *C4.5* est le plus performant avec un taux de précision (*Acc*) égal à 89.50%.

La méthode de *sur-échantillonnage* permet de rétablir un certain équilibre au niveau de l'ensemble des données; cela consiste à accroître la proportion d'instances appartenant à la classe minoritaire par une reproduction synthétique d'instances effectuée de manière aléatoire. Par contre, la méthode de *sous-échantillonnage* implique la suppression aléatoire d'instances de la classe majoritaire, de sorte à obtenir une répartition plus équilibrée des classes. En ce sens, le ré-échantillonnage des données a pour but de prévenir ce problème de déséquilibre

des données de manière à rétablir une situation plus correcte/normale, ainsi que d'améliorer l'apprentissage en vue d'obtenir un modèle plus robuste.

Relativement peu de méthodes traitent le problème lié au déséquilibre des classes au niveau des algorithmes consistant à faire adapter les modèles d'apprentissage traditionnels/existants de manière à atténuer le biais contre les classes majoritaires tout en les ajustant aux données minières présentant des distributions biaisées [144]. Une telle branche n'est pas aussi répandue parmi les chercheurs en raison de sa complexité en termes de conception, et qui dépend directement de l'ensemble de données utilisé.

Par ailleurs, l'apprentissage d'ensemble est de plus en plus répandu en tant que moyen de surmonter les problèmes relatifs au déséquilibre des classes [144]. A ce titre, l'apprentissage d'ensemble (combinaison de classifieurs) représente la tendance largement adoptée dans le cadre des systèmes d'aide au diagnostic médical. Actuellement, la combinaison de classifieurs occupe une place importante dans le domaine de l'apprentissage automatique et de l'intelligence artificielle en raison de ces résultats honorables obtenus pour une variété d'applications [8, 65, 110, 145].

Sarwar et al. [146] ont présenté une méthode d'ensemble reposant sur le vote au moyen de quinze algorithmes d'apprentissage automatique pour le diagnostic du diabète type 2. Dans le cadre de cette étude, seulement quatre paradigmes (K-ppv, SVM, RNA et NB) sont pris en compte, constituant ainsi le pool de classifieurs de base, tandis que le vote majoritaire est adopté afin de parvenir à une décision finale présentant une précision de 98.60%. Maniruzzaman et al. [147] ont opté en faveur d'une analyse de plusieurs classifieurs dont le naîf de bayes (NB), l'arbre de décision (AD), les forêts aléatoires (RF) et AdaBoost (AB) dans le but de maintenir le classifieur le plus efficace permettant de distinguer les patients normaux des diabétiques. L'algorithme de régression logistique (*RL*) est considéré dans ce travail en tant que technique de sélection des caractéristiques pour déterminer les facteurs de risque du diabète en fonction de la valeur p et du rapport de cotes (*OR*). Le paradigme RL basé sur le classifieur d'ensemble *forêts aléatoires* a obtenu les meilleurs résultats avec une précision de 94.25% en appliquant le protocole de validation croisée à 10 $blocs$.
Mahabub [148] a proposé de mettre en œuvre un système de détection précoce du diabète à travers une étude analytique de plusieurs paradigmes d'apprentissage automatique tels que K-ppv, AB, AD, RF, SVM, RL, PMC, GradientBoosting (GB), MultinomialNB, ExtremeGB et GaussianNB avec pour objectif de conserver uniquement les trois classifieurs les plus

efficaces (K-ppv, SVM et PMC) et de les réutiliser ensuite dans la construction du pool de classifieurs de base en adoptant une approche de vote (vote hard et vote soft). Sur la base des résultats expérimentaux menés dans le cadre de ce travail, la méthode de vote d'ensemble surpasse tous les autres classifieurs pris séparément avec une précision de 86.00% en utilisant la technique de validation croisée à 10 $blocs$.

Hasan et al. [149] ont examiné plusieurs méthodes d'ensemble basées sur le vote ainsi que différentes combinaisons de divers classifieurs de base (tels que k-ppv, AD, RF, AB, NB et XGBoost (XB)) en vue de sélectionner le pool le plus robuste en utilisant le vote pondéré soft conformément aux valeurs AUROC pour stimuler la prédiction du diabète. Toujours dans le cadre de cette étude, les auteurs comparent la performance de l'algorithme PMC (en tant que classifieur unique) à celle d'autres schémas envisagés. Une étape préliminaire comprenant le rejet des valeurs aberrantes, le remplissage des valeurs manquantes/nulles, la standardisation des données (*normalisation du score z*) et la sélection des caractéristiques (*ACP, ICA et approche basée sur la corrélation*) fait partie de la méthode suggérée. En effet, selon les résultats expérimentaux, le pool composé des classifieurs de base *AB & XB* a atteint des performances supérieures à celles des autres combinaisons envisageables ainsi qu'à l'algorithme PMC, ce qui a permis d'obtenir un AUROC de 95.00% en employant le protocole de validation croisée à 5 $blocs$.

L'idée principale au sujet de l'apprentissage d'ensemble contribue à améliorer la performance de l'apprentissage déséquilibré en fusionnant divers classifieurs de manière à diminuer la variance, le biais, et/ou à renforcer les prédictions. En matière d'apprentissage automatique et d'intelligence artificielle, les stratégies d'apprentissage d'ensemble consistent à tirer parti de l'expertise variée de différents classifieurs pour obtenir, grâce à la combinaison, un modèle de décision finale censé être meilleur que tout autre classifieur lorsqu'il est utilisé individuellement/séparément.

La figure 4.2 ci-dessous représente ce procédé pour un problème à deux classes linéairement séparables. Les points bleus décrivent les échantillons de la première classe et les triangles rouges ceux de la seconde classe. Le succès en matière de méthodes d'apprentissage d'ensemble réside dans le fait de garantir un taux d'erreur moins important que celui fourni par son meilleur classifieur.

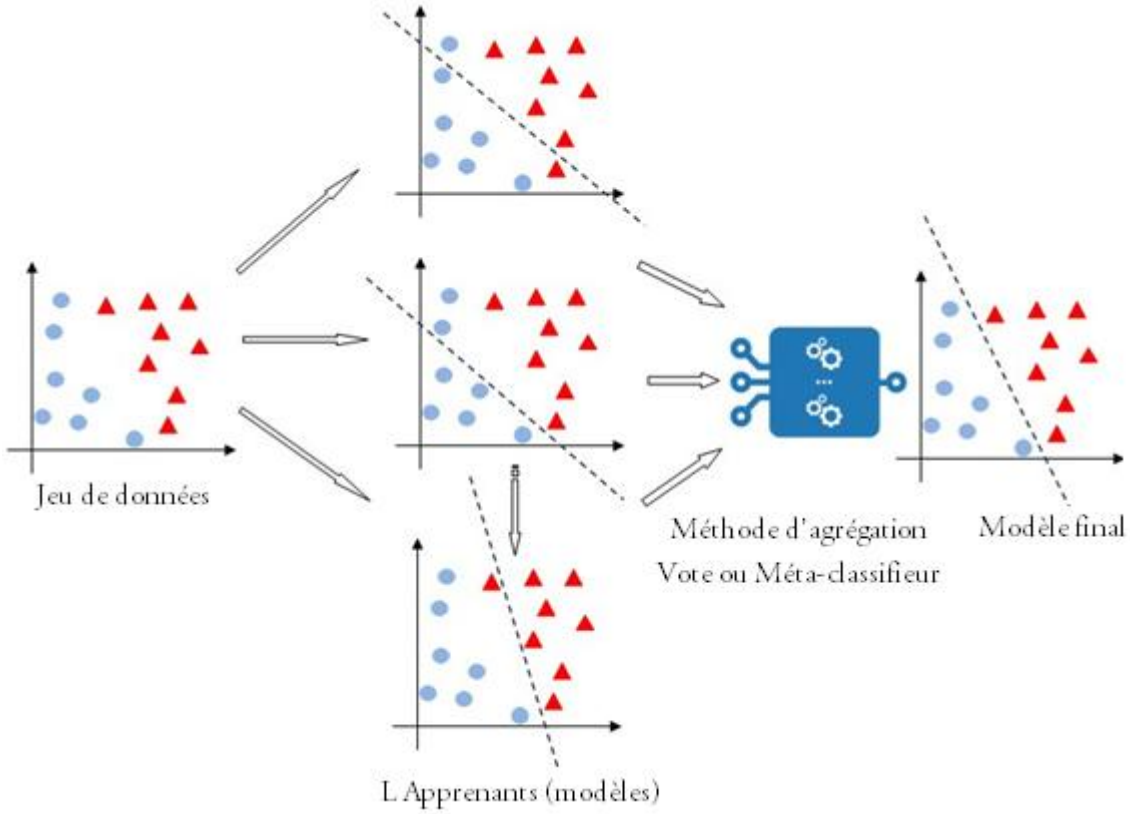

Figure 4.2: Processus général relatif aux méthodes d'apprentissage d'ensemble

Stacking (également connu sous le nom de *méta-classification*) constitue un paradigme d'apprentissage d'ensemble permettant de combiner plusieurs classifieurs de base par le biais d'un méta-classifieur de haut niveau, et ce, à la différence d'autres algorithmes basés sur le vote, tels que le bagging et le boosting. En effet, la combinaison des caractéristiques (*features*) propres à plusieurs modèles de classification en vue de créer un nouveau modèle plus performant. De plus, la variabilité ainsi que la diversité existant entre divers classifieurs composant le modèle représentent en réalité un facteur important afin de tirer avantage de la complémentarité pouvant éventuellement être présente entre eux. Cela revient à générer un ensemble de classifieurs hétérogènes/complémentaires (*multimodaux*) dont la combinaison (*méta-apprentissage*) permet de parvenir à une solution optimale.

Cette diversité/complémentarité constitue une des propriétés fondamentales associées au concept de *multimodalité*.

Il importe également dans ce contexte de souligner que la majorité des études mentionnées dans la littérature [139-156] se concentrent le plus souvent autour d'un classifieur unique relatif à la phase de diagnostic/classification du diabète, tandis que peu de travaux recourent aux algorithmes d'apprentissage d'ensemble basés principalement sur des classifieurs de base

homogènes ainsi que sur des techniques de vote impliquant notamment une agrégation au niveau de classe, fournissant ainsi moins d'informations dans le cadre de la fusion; alors que très peu d'études traitent de la notion de déséquilibre des données au niveau des données au moyen de la technique *SMOTE*.

La présente contribution aborde de manière hybride l'apprentissage déséquilibré en proposant une méthode de méta-classification multimodale améliorée appelée *IRESAMPLE+St* sur la base du paradigme Stacking tout en utilisant la complémentarité susceptible d'exister entre les différents classifieurs en vue de faire la distinction entre les patients sains par rapport aux patients diabétiques. Parallèlement, cette étude met l'accent sur une stratégie modifiée dérivée de la méthode RESAMPLE, désignée sous le nom *d'IRESAMPLE+*, ainsi que sur la méthode SMOTE, qui sont intégrées à titre d'étape préliminaire dans le processus de ré-échantillonnage en vue de surmonter et de résoudre le problème lié aux données déséquilibrées; et ceci afin de former les classifieurs de base du pool ainsi que pour obtenir un modèle plus efficace.

Ce pool se compose essentiellement de cinq classifieurs de base hétérogènes, à savoir le perceptron multicouches (*PMC*), les k-plus proches voisins (*K-ppv*), la machine à vecteurs de support (*SVM*), les forêts aléatoires (*RF*) et le naïf de bayes (*NB*). Le modèle de classification en question est créé sur la base de divers types de classifieurs (de base) destinés à être utilisés sous forme de méta-caractéristiques (*meta-features*) en tant que nouvelles données d'entrée au niveau d'un méta-classifieur, en abordant la fusion au niveau des scores à travers des informations importantes disponibles à ce niveau. La tâche principale du méta-classifieur ou de l'agrégateur stacking, désigné par l'algorithme *SVM*, consiste alors à apprendre la meilleure façon de combiner les différents modèles de classification de manière à obtenir le classifieur le plus précis/performant.

Dans cette étude, les points de recherche mentionnés ci-après font l'objet d'un examen approfondi:

- Le prétraitement des données joue-t-il un rôle important dans la performance du système proposé?
- Quelle technique de ré-échantillonnage est la mieux adaptée pour équilibrer les données?
- Quelle méthode d'agrégation convient le mieux à la combinaison de classifieurs?
- La combinaison de classifieurs a-t-elle un impact par rapport à la performance?

Les principaux apports du présent chapitre sont mis en évidence par les aspects suivants:

- Hybridation entre deux approches de classification des données déséquilibrées, à savoir celle basée au niveau des données et celle fondée au niveau des algorithmes.
- Proposition d'une stratégie de ré-échantillonnage modifiée appelée $IRESAMPLE+$ basée sur RESAMPLE qui fonctionne avec succès à la fois comme technique de sur-échantillonnage et de sous-échantillonnage en vue d'améliorer les performances de classification sur l'ensemble de données médicales déséquilibrées *Pima Indians Diabetes (PID)*.
- Analyser le comportement en matière d'apprentissage d'ensemble sur la base d'une méta-classification, connu sous le nom *d'agrégateur stacking*, tout en intégrant un module d'apprentissage croisé « *cross-training* » ainsi qu'en générant un ensemble de classifieurs de base complémentaires.

Le suivi de ce chapitre illustre en premier temps l'architecture globale de l'approche proposée, puis une présentation détaillée des principaux axes du système proposé $IRESAMPLE+St$ dans le cadre du diagnostic du glaucome ainsi que les résultats expérimentaux de ce travail, et se termine également avec une conclusion.

2 Système proposé IRESAMPLE+St

La méthodologie relative au système proposé IRESAMPLE+St se présente à travers un schéma illustré dans la figure 4.3. La principale préoccupation exprimée dans cette étude consiste à mettre en équilibre les données grâce à l'utilisation de la technique RESAMPLE améliorée, appelée $IRESAMPLE+$; en outre, cette étude vise également à produire un ensemble de classifieurs complémentaires de même que l'analyse de différentes approches de fusion par le biais d'une nouvelle stratégie de fusion dite de méta-classification.

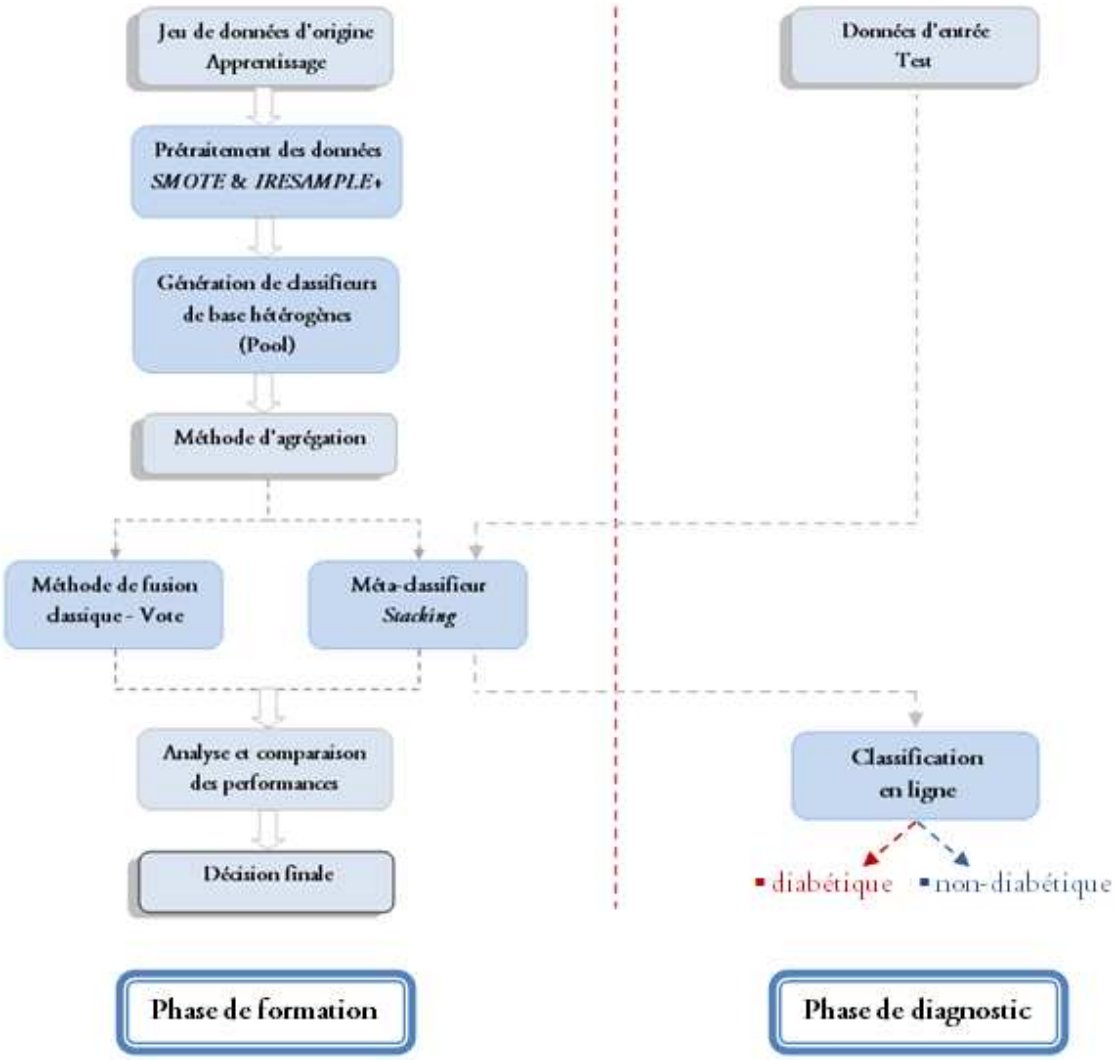

Figure 4.3: Architecture de l'approche suggérée concernant le diagnostic du diabète

2.1 Prétraitement des données

La phase de prétraitement des données est perçue comme étant nécessaire dans le domaine de l'apprentissage automatique et de la fouille de données dans le but d'éliminer les données bruitées tout en formant les classifieurs sur une meilleure qualité de données. Le ré-échantillonnage des données est une notion fréquemment utilisée afin de diminuer l'impact du déséquilibre des données dans le cadre de l'apprentissage. En règle générale, le concept de ré-échantillonnage consiste à réduire celui de la classe majoritaire (*sous-échantillonnage*), ou à accroître celui de la classe minoritaire (*sur-échantillonnage*). Ainsi, pour un problème de classification comportant deux classes, on parle d'une classe minoritaire disposant d'un nombre relativement inférieur d'échantillons par rapport à la classe majoritaire qui en compte un nombre proportionnellement plus important.

Dans le cadre de cette étude, l'étape préliminaire pour le prétraitement des données consiste en la mise en œuvre des techniques appropriées en matière de ré-échantillonnage. Il s'agit de constituer un ensemble de données équilibré permettant une représentation/ distribution des données optimale afin de former le pool de classifieurs de base (ce qui favorise un apprentissage équilibré), et ce, au moyen de la technique du SMOTE ainsi que de la méthode modifiée reposant sur la notion de RESAMPLE, appelée $IRESAMPLE +$, communément utilisées compte tenu de ses meilleures performances dans le domaine de l'apprentissage automatique [157].

De manière générale, la technique SMOTE (stratégie de sur-échantillonnage supervisé) consiste en la production d'instances artificielles permettant d'étendre les limites de la classe minoritaire; celles-ci sont générées aléatoirement le long des segments de ligne d'un certain nombre de k-plus proches voisins appartenant à la même classe. Ainsi, cette stratégie rend la zone de classe minoritaire plus grande et générale. La stratégie *SMOTE* repose sur quatre paramètres afin de déterminer le nombre d'instances devant être créées:

$$SMOTE \leftarrow fonction\ (S, P, K, C) \tag{4.1}$$

Où,

$-S$ indique la graine de nombre aléatoire, $-P$ désigne le pourcentage d'instances *SMOTE* à établir, $-K$ est le nombre de plus proches voisins à utiliser, et $-C$ définit l'indice de la valeur de la classe nominale de *SMOTE* (classe minoritaire). Un exemple de génération d'instances synthétiques est présenté ci-dessous (Figure 4.4).

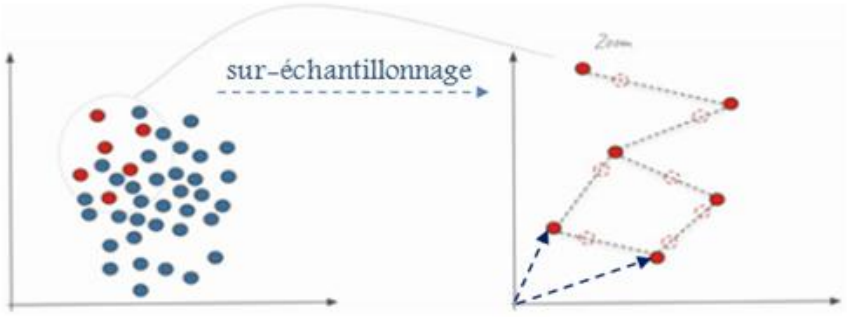

Figure 4.4: Exemple illustrant le principe de SMOTE [157]

Ce principe est basé sur la différence entre une instance minoritaire et son plus proche voisin. Cette différence est multipliée par un nombre aléatoire compris entre 0 et 1, puis ajoutée au vecteur caractéristique. Le pseudo-code qui suit décrit ce concept:

Pseudo-code: SMOTE (stratégie de sur-échantillonnage supervisé)

Entrée: jeu de données déséquilibré (*diabète_PID*)

Sortie: ensemble de données équilibré

Processus:

1. Pour un nombre suffisant d'instances synthétiques, faire
2. Sélection d'une instance minoritaire I
3. Choix de l'une des instances voisines les plus proches N
4. Sélection d'un poids aléatoire entre 0 et 1 W
5. Création de la nouvelle instance synthétique S
6. Pour chaque attribut, faire
7. Calcul: $valeurS = valeurI + (valeurN - ValeurI) \times W$
8. Fin pour
9. Fin pour

Dans le cadre de la technique RESAMPLE (stratégie de sous-échantillonnage supervisé), qui consiste à créer un sous échantillon randomisé d'un ensemble de données en procédant à un échantillonnage avec ou sans remplacement, dans la mesure où chaque instance du sous-ensemble présente la même probabilité d'être choisie. Les termes suivants désignent le nombre d'instances à sélectionner dans le cadre de la stratégie *RESAMPLE*:

$$RESAMPLE \leftarrow fonction\ (S, Z, B, no - replacement, V) \tag{4.2}$$

Où,

$-S$ représente la graine de nombre aléatoire, $-Z$ indique la taille du jeu de données de sortie sous forme de pourcentage du jeu de données d'entrée, $-B$ correspond au facteur de biais vers une distribution uniforme des classes, $-no - replacement$ désactive le remplacement des instances, et $-V$ renverse le choix (disponible uniquement avec $-no - replacement$ $-$). La figure 4.5 montre un exemple de la méthode de *RESAMPLE*.

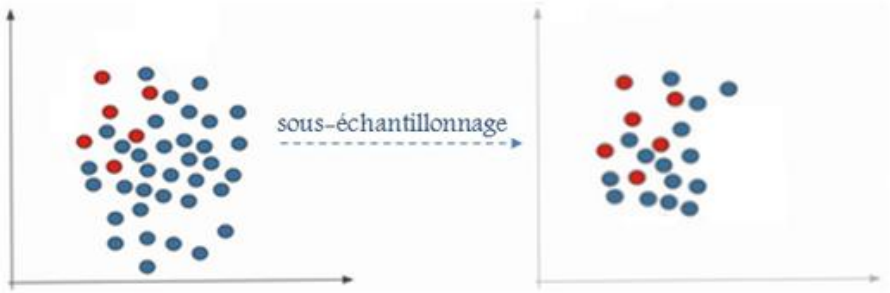

Figure 4.5: Exemple illustrant le principe de RESAMPLE [157]

Le but étant en effet d'analyser chacune de ces deux méthodes (*SMOTE* et $IRESAMPLE +$ -basée sur la technique *RESAMPLE*-) de manière à conserver la stratégie la plus efficace.

2.2 Approche de classification d'ensemble: classifieurs de base et stratégie d'agrégation

Dans la mesure où aucun modèle de classification uniforme ne permet de résoudre simultanément et de manière exhaustive tout type de problème d'apprentissage automatique, il est également impossible de disposer de classifieur optimal en mesure d'apprendre toute répartition de données d'apprentissage. Cela explique pourquoi il est difficile de faire ressortir l'excellence d'un paradigme d'apprentissage automatique au détriment d'un autre; de ce fait, il convient de mettre l'accent sur le recours à la combinaison de classifieurs.

En effet, l'objectif de la combinaison de classifieurs réside dans la possibilité de construire simultanément un ensemble de classifieurs divers/complémentaires et efficaces dans le but de parvenir à une meilleure prise de décision au moyen de diverses stratégies en matière d'agrégation.

La méthode d'apprentissage multimodale $IRESAMPLE + St$ se base en effet sur le principe de la fusion précoce [1], ce qui signifie concrètement que les différentes sorties/caractéristiques en provenance des classifieurs de base seront combinées entre elles « *meta-features* » constituant ainsi des entrées relatives à un second modèle de classification « *méta-classifieur* » en vue de générer le résultat final. Ainsi, aucune décision concernant le diagnostic du diabète ne peut être prise dès lors que les caractéristiques/sorties de divers classifieurs sont fusionnées.

Dans le cas présent, le rôle des classifieurs de base consiste essentiellement en différentes modalités, où ces différents classifieurs apportent certains types d'informations spécifiques ainsi qu'un point de vue propre/unique et ce, d'une part, en tirant parti de l'indépendance entre les classifieurs de base. D'autre part, en profitant de la complémentarité/diversité susceptible de se présenter dans le cadre de ces classifieurs.

Cette étude adopte le concept d'ensemble parallèle consistant en la production en parallèle de "m" classifieurs de base distincts. Cette approche nous a permis de créer un ensemble de classifieurs de base au terme de plusieurs expériences et après une étude détaillée de la

littérature relative à d'autres tâches d'apprentissage automatique. Le schéma ci-dessous (Figure 4.6) montre la structure du système proposé pour le diagnostic du diabète.

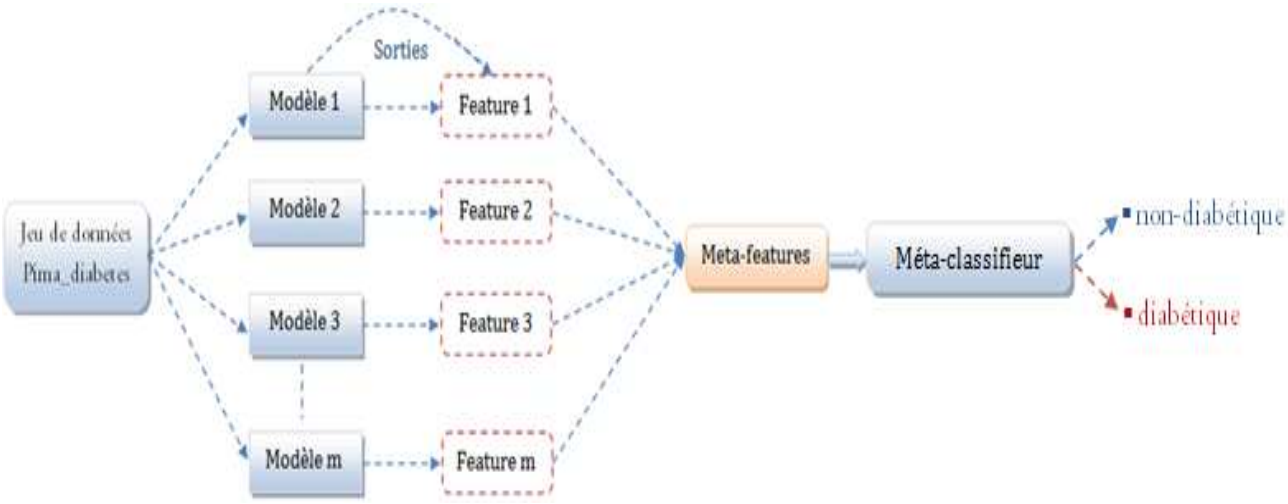

Figure 4.6: Conception de réseau proposé *IRESAMPLE+St* en vue de diagnostiquer le diabète

Il convient de procéder à une approche d'agrégation appropriée après avoir formé les apprenants/classifieurs de base de manière à rassembler leurs sorties sous une forme unique destinée à servir par la suite à la classification finale. Diverses approches sont proposées permettant de combiner différents classifieurs de base (complémentaires) du pool, comme le combineur linéaire, le combineur produit et le combineur vote.

Dans le but de parvenir à la décision finale au sujet de l'approche d'ensemble proposée, il est procédé à la mise en œuvre de deux méthodes avancées de fusion de type classe (dont *le vote à la majorité (VM)* et *le vote à la majorité pondérée (VMP)*), de même que la stratégie de *l'agrégateur stacking* ou «*méta-classifieur*».

Dans le cadre de la première stratégie de fusion basée sur le vote, les règles de combinaison VM et VMP font partie des techniques les plus fiables à travers plusieurs domaines d'application [1, 65, 146, 148, 149]. Le principe de telles techniques est déjà élaboré au chapitre 3 (voir Section 1.3).

En ce qui concerne la deuxième stratégie de fusion utilisée au cours de cette étude, à savoir le méta-classifieur reposant sur le paradigme du Stacking. Cette méthode d'apprentissage d'ensemble consiste en un recours à des classifieurs qui, souvent, sont hétérogènes; autrement dit, qui fait appel à des apprenants de types distincts, menant ainsi à un ensemble hétérogène. Dans son principe général, cet algorithme vise à apprendre à combiner de manière optimale

les prédictions de plusieurs modèles d'apprentissage machine performants au moyen d'un méta-classifieur.

L'architecture d'un modèle de Stacking implique deux modèles de base ou plus, souvent appelés modèles de premier niveau, et un méta-modèle qui combine les prédictions des modèles de base, appelé modèle de deuxième niveau. Les modèles de base se forment sur la base d'un ensemble de données de formation initiale, et ensuite, le *méta-classifieur* ou le méta-modèle se forme autour des sorties/prédictions des modèles de base à titre de caractéristiques « *méta-features* ».

En effet, deux arguments sont à définir dans le cadre de la construction du modèle d'ensemble proposé: les apprenants de base "m" à adapter et le méta-apprenant responsable de la mise en place de l'agrégation. Ces apprenants de base ($m = 5$) correspondent aux algorithmes suivants, soit le perceptron multicouches (*PMC*), les k-plus proches voisins (*K-ppv*), la machine à vecteurs de support (*SVM*), les forêts aléatoires (*RF*) et le naïf de bayes (*NB*). Au stade de la classification/diagnostic, le méta-apprenant *SVM* (en prenant en compte les caractéristiques propres aux cinq apprenants de base) est adopté grâce à sa performance face aux problèmes populaires liés à la classification binaire qui est considérée dans le domaine du diagnostic médical comme étant le meilleur séparateur binaire.

L'approche d'ensemble proposée reposant sur la méthode de *Stacking* intègre la stratégie de $formation\ croisée\ k-blocs$ « *k-fold cross-training* » pour $k = 10$ (qui est semblable à la technique de validation croisée) en vue de construire le méta-modèle, de sorte que toutes les instances soient utilisées pour former le méta-classifieur et le modèle final. Ce procédé consiste à découper au hasard (sans remplacement) l'ensemble de données d'origine en k échantillons équivalents ou k parties de taille approximativement égale. La première partie est conservée pour les tests et le modèle est entraîné sur les $k - 1$ parties.

Cela signifie que les apprenants de base sont alors entraînés sur les $k - 1$ parties et sont validés/testés au moyen d'une des k parties restantes de sorte que ce processus soit répété k fois, de manière à ce que chaque k sous-partie soit utilisée exactement une fois en tant qu'ensemble de validation. De cette façon, le principe de l'opération de Stacking « *méta-classification* » se traduit à travers l'algorithme suivant en utilisant la méthode de $formation\ croisée\ k-blocs$.

Algorithme: Méta-classification en utilisant un apprentissage croisé k *blocs*

Entrée: formation d'un ensemble de données $\mathcal{D} = \{x_i, y_i\}\ i \leftarrow 1$ à n

algorithmes d'apprentissage de premier niveau $C1, \ldots, CT$

algorithme de méta-apprentissage de second niveau C'

Sortie: un ensemble de classifieurs C

Processus:

1. Étape 1: utilisation de la technique de validation croisée en vue de préparer un ensemble de données pour l'apprentissage du méta-classifieur
2. Répartition aléatoire de $\mathcal{D}$ en K sous-parties de taille uniforme: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \ldots, \mathcal{D}_k\}$
3. Pour $k \leftarrow 1$ à K faire
4. Étape 1.1: apprentissage des classifieurs de base
5. Pour $t \leftarrow 1$ à T faire
6. apprentissage d'un classifieur C_{kt} à partir de $\mathcal{D} \backslash \mathcal{D}_k$
7. Fin pour
8. Étape 1.2: Construction d'un ensemble de données pour l'apprentissage du méta-classifieur
9. Pour $x_i \in \mathcal{D}_k$ faire
10. $D_C = \{x'_i, y_i\}$, où $x'_i = \{C_{k1}(x_i), C_{k2}(x_i), C_{k3}(x_i), \ldots, C_{kT}(x_i)\}$
11. Fin pour
12. Fin pour
13. Étape 2: apprentissage d'un méta-classifieur
14. Apprentissage d'un nouveau classifieur C' sur la base de D_C
15. Retourner $C(x) = C'(C_1(x), C_2(x), C_3(x), \ldots, C_T(x))$

2.3 Complexité de calcul

La complexité de calcul au niveau des algorithmes représente un sujet d'étude intéressant dans le domaine de l'apprentissage automatique. En règle générale, la complexité d'un modèle de l'apprentissage automatique s'évalue par le biais de la notation $Big - O$. Ce concept notamment deux volets : la complexité de temps, qui se rapporte au temps d'exécution du modèle/algorithme (il est toujours spécifié en termes relatifs à une certaine taille d'entrée "n"),

ainsi que la complexité mémoire, relative à la quantité de mémoire consommée par le ce dernier.

La méthode d'ensemble de méta-classification proposée se compose d'un pool de classifieurs de base "m", d'un certain nombre d'échantillons d'apprentissage "n" de même qu'un ensemble de caractéristiques "f" généré au moyen des classifieurs de base à titre de données d'entrée destinées au méta-classifieur. Il convient de mentionner que le concept de coût de calcul ne s'applique pas au processus de classification en ligne (*en temps réel*), mais uniquement à la phase d'apprentissage (*hors ligne*) dont le temps de calcul est faible et raisonnable, soit environ 15 $minutes$.

Cette approche $IRESAMPLE + St$ se définit ainsi au regard de sa complexité de calcul: complexité_temps_apprentissage $= O(St(n^2.m.f))$ et complexité_mémoire $= O(m.f)$. Où "m" indique la valeur de la somme relative à la complexité temporelle (en termes de notation $Big - O$) associée à tous les classifieurs utilisés (*PMC*, *K-ppv*, *SVM*, *RF* et *NB*), sachant que PMC correspond à $O(n^2.f.l_i)$, K-ppv vaut $O(n.f.k_{\text{nombre de voisins}})$, SVM équivaut à $O(n_{sv}^2.f)$, RF égale $O(n.log(n).f.t_{nombre\ d'arbres})$, et NB étant $O(n.f)$. A ce propos, sv représente le nombre de vecteurs de support, l est égal au nombre de couches cachées, et i indique le nombre de neurones dans chaque couche.

3 Évaluation des résultats expérimentaux

En vue d'évaluer la méthode de fusion « méta-classifieur » aux fins du diagnostic du diabète, la technique de validation croisée $k - blocs$ est appliquée avec $k = 10$. La mise en œuvre de la tâche proposée est effectuée par le biais de l'outil Weka, librairie libre dédiée aux algorithmes d'apprentissage automatique, écrite en Java, et donnant également l'accès à des outils réputés tels que *deeplearning4j*, *scikit-learn* et *R*. Les subdivisions suivantes exposent en détail les expériences menées dans le cadre de ce travail ainsi que les résultats achevés.

3.1 Description de l'ensemble de données sur lequel porte cette étude

Dans le cadre de la méthode proposée en matière de diagnostic du diabète, un ensemble de 768 instances et de 8 attributs de la base de données *Pima Indians Diabetes (PID)*[5] est utilisé,

[5] https://www.kaggle.com/

dont 500 cas correspondent aux patients en bonne santé par rapport à 268 cas souffrant de diabète. Ces attributs du *PID* sont décrits dans le tableau 4.1.

Tableau 4.1: Brève description au sujet des attributs de la base de données du diabète (PID)

Nom de l'attribut	Description	Valeur		
		Min	Max	Moyenne
Pregnancies	Nombre de fois enceinte	0	17	3.845
Glucose	Concentration de glucose plasmatique a 2 heures lors d'un test oral de tolérance au glucose	0	199	120.895
BloodPressure	Pression sanguine diastolique $(mm\ Hg)$	0	122	69.105
SkinThickness	Épaisseur du pli cutané du triceps (mm)	0	99	20.536
Insulin	Insuline sérique 2 heures $(mu\ U/ml)$	0	846	79.799
BMI	indice de masse corporelle IMC $(poids\ en\ kg/(taille\ en\ m)\text{^}2)$	0	67.1	31.993
DiabetesPedigreeFunction	Fonction généalogique du diabète	0.08	2.42	0.472
Age	Âge (années)	21	81	33.241
Outcome	Variable de classe (négative ou positive)	0	1	-

3.2 Critères d'évaluation

Dans le but d'évaluer la performance relative au système proposé *IRESAMPLE+St*, il convient à cet effet de se référer aux mesures de performance en usage courant dans le domaine du diagnostic médical, telles que la précision (Acc), la sensibilité (Sen), la spécificité (Spe), la valeur prédictive positive (VPP), la valeur prédictive négative (VPN) et le coefficient de corrélation Matthews (CCM) qui sont présentées précédemment respectivement au moyen des équations (3.8), (3.9), (3.10), (3.11) et (3.12). De même, le critère $F-score$ fait également partie de ces dernières mesures utilisées dans ce travail qui est défini par l'équation (4.1):

$$F-measure\ (F-score)\ = 2\ .\ \frac{VPP\ .\ Sen}{VPP+Sen} \tag{4.1}$$

Où, les vrais positifs (VP), les vrais négatifs (VN), les faux positifs (FP) de même que les faux négatifs (FN) sont indiqués ci-dessous dans le tableau 4.2.

Tableau 4.2: Matrice de confusion liée à la classification binaire (diabète)

	Test positif (P)	**Test négatif (N)**
Patients diabétiques (P)	VP	FN
Patients non diabétiques (N)	FP	VN

En outre, cette méthode proposée est également mesurée selon l'aire sous la courbe *AUROC* (*receiver operating characteristic - caractéristique d'efficacité du récepteur*) [51].

3.3 Résultats et discussion

Diverses expériences portant sur la constitution du pool de classifieurs de base et plus précisément sur le nombre de classifieurs choisis sont menées au cours de cette étude [158, 159], en adoptant cinq modèles distincts qui sont les suivants: le perceptron multicouches (*PMC*), les k-plus proches voisins (*K-ppv*), la machine à vecteurs de support (*SVM*), les forêts aléatoires (*RF*) et le naïf de bayes (*NB*). De tels classifieurs sont par ailleurs formés à partir de l'ensemble des données originales, c'est-à-dire sans aucun traitement préalable, pour être en mesure de mieux évaluer et de comparer les résultats obtenus lors de ces expérimentations.

3.3.1 Mise en place expérimentale

La configuration des paramètres ainsi que les performances de chaque algorithme du système suggéré sont détaillées dans les paragraphes suivants:

Tout comme pour le classifieur PMC, les résultats idéaux sont obtenus par le biais du réglage de certains paramètres : "$couches\_cachées: 'a' = ((attribs + classes)/2) = 5$", "$taux\_d'apprentissage = 0.3$", "$momentum = 0.2$", "$nombre\ d'époques = 500$", et "$seuil\_de\_validation = 20$". Les performances pertinentes obtenues sont les suivantes: un Acc de 75.39%, une Sen de 60.80%, une Spe de 83.20%, un F-score de 63.29%, un CCM de 44.92% et un AUROC de 79.30%.

Concernant le classifieur K-ppv, les paramètres relatifs au nombre K des voisins à utiliser, à l'algorithme de recherche du plus proche voisin, à la pondération de la distance ainsi qu'à la taille de la fenêtre sont fixés de manière à obtenir des résultats optimisés, à savoir "$K = 2$", "$Algorithme$": $Distance\ Euclidienne - R\ premier - dernier$", "$Pondération\_de\_la\_distance: Non$", et "$taille\_de\_la\_fenêtre = 0$". Les valeurs

correspondantes ainsi obtenues sont celles de 72.66% Acc, 55.20% Sen, 82.00% Spe, 58.49% F-score, 38.37% CCM et 74.20% AUROC.

Les résultats optimaux sont atteints par rapport au classifieur SVM en mettant en place les principales variables suivantes: noyau, gamma et C, où "$noyau$": $Polynôme$" avec une valeur $gamma$ de 0.5 et un paramètre C égal à 3. Les résultats correspondants obtenus sont: Acc 77.47%, Sen 55.60%, Spe 89.20%, F-score 63.27%, CCM 48.42% et AUROC 72.40%.

Les paramètres à régler ainsi que leurs résultats optimisés qui sont obtenus à l'aide du classifieur RF se déterminent comme suit: "$profondeur\_maximale$: $illimitée$", "$nombre\_Features = (< 0 = int(log\_2(\#prédicteurs) + 1))$", et "$nombre\_Arbres = 100$". Les résultats correspondants atteints comprennent: Acc 74.87%, Sen 59.00%, Spe 83.40%, F-score 62.10%, CCM 43.51% et AUROC 81.50%.

Le classifieur NB est plus efficace lorsque les paramètres $displayModelInOldFormat$ et $useKernelEstimator$ sont réglés en mode $faux$ et le paramètre $useSupervisedDiscretization$ est défini à $vrai$. Les performances correspondantes représentent un Acc de 76.30%, une Sen de 61.20%, une Spe de 84.40%, un F-score de 64.30%, un CCM de 46.78% et un AUROC de 81.90%.

Les expérimentations de ce travail sont menées avec un système d'exploitation Windows-7 avec la configuration matérielle suivante: Processeur *Intel® Core™ i7-8750H à 2,20 GHz jusqu'à 4,10 GHz* avec *RAM 64 Go* et mémoire *DDR4-2666, LPDDR3-2133*.

3.3.2 Mise en équilibre de l'ensemble des données via la stratégie de sur- et sous-échantillonnage « IRESAMPLE+ »

La présente étude examine la manière d'utiliser le pool de classifieurs de base dans le cadre des données prétraitées/équilibrées de la base de données PID. Autrement dit, cette méthode suggérée est basée sur les résultats obtenus grâce aux techniques d'équilibrage des données (*SMOTE*, *RESAMPLE* et *IRESAMPLE+*). À titre d'illustration de ce recours aux techniques de ré-échantillonnage concernant l'approche proposée, les différentes phases à suivre pour équilibrer les données en utilisant le filtre *SMOTE* et la méthode modifiée basée sur RESAMPLE appelée *IRESAMPLE+* sont ainsi présentées ci-après.

Il est à noter que la figure 4.7 montre qu'il y a une grande non-uniformité au niveau de la distribution des données entre les classes du jeu de données (sur le diabète_*PID*) utilisé dans ce travail.

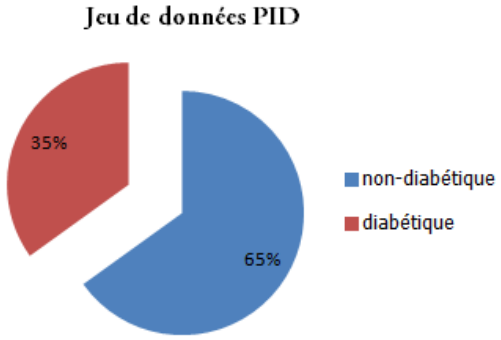

Figure 4.7: Déséquilibre de classes de la base de données du diabète (PID)

La disparité entre les classes finales telles que la classe négative (*non-diabétique*) et la classe positive (*diabétique*) peut aboutir à des résultats de diagnostic incorrects. Ainsi, le classifieur est susceptible d'être davantage biaisé vers la classe à forte concentration (*non-diabétiques*) que vers la classe à faible concentration (*diabétiques*), ce qui peut également conduire à une classification inexacte des patients pour détecter la catégorie correspondante.

Tenant compte de la nature de déséquilibre de la classe de sortie, à savoir $testée\_positive$ (classe diabétique), une stratégie de sur-échantillonnage, *SMOTE*, est appliquée à l'ensemble de données relatives au diabète. La stratégie de filtrage supervisé (*SMOTE*) est mise en œuvre en utilisant les paramètres de l'équation (4.2) comme suit:

$$SMOTE \leftarrow fonction\ (1, 86.6, 5, 2) \tag{4.2}$$

L'utilisation de la technique SMOTE (*avec cette configuration*) a permis de faire passer la taille de la classe minoritaire (*diabétiques*) de 268 à 500 afin de disposer ainsi d'une base de données équilibrée tout en obtenant des résultats significatifs. La figure 4.8 présente un schéma comparatif entre l'ensemble des données brutes contre l'application de la méthode *SMOTE*.

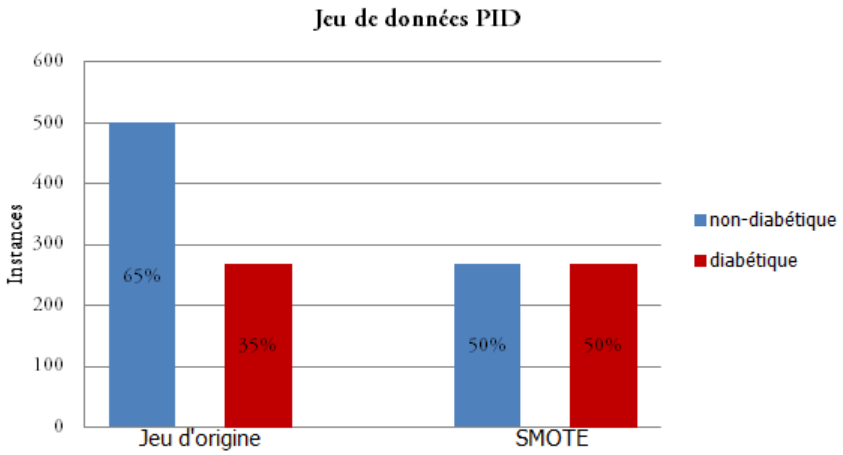

Figure 4.8: Application du filtre SMOTE en matière de distribution des classes

Par ailleurs, la stratégie de sous-échantillonnage, *RESAMPLE*, est également pratiquée auprès de la classe majoritaire (non diabétique) de manière à réduire la taille de celle-ci de 500 à 268, de sorte que les deux classes possèdent un nombre identique d'instances, et cette stratégie de filtrage supervisé (*RESAMPLE*) est configurable de la manière suivante:

$$RESAMPLE \leftarrow fonction\ (1, 69.8, 1, no - replacement, F) \tag{4.3}$$

Une telle configuration a engendré en effet une grande perte au niveau des données (instances), dont les résultats sont insatisfaisants. La présente étude vise à modifier/améliorer le fonctionnement général et le principe de cette technique, de manière à ce que celle-ci puisse opérer à la fois en tant que stratégie de sur- et de sous-échantillonnage et ce, à travers un ajustement approprié au niveau des paramètres de l'équation (4.3) tout en introduisant également une structure itérative permettant de procéder à l'équilibrage significatif des données - le pseudo-code qui suit permet de mettre en évidence cet aspect:

Pseudo-code: Processus *IRESAMPLE+* (stratégie de sur- et de sous-échantillonnage)

Entrée: Ensemble de données original non équilibré (diabète_PID_)

Sortie: Ensemble de données équilibré plus efficace

Processus:

1. Début
2. Initialisation
3. Répéter
4. $RESAMPLE \leftarrow fonction\ (1, 100, 0, no - replacement, F)$ (4.4)
5. Jusqu'à (nombre d'instances de classe $1_{\text{non-diabétiques}}$ = nombre d'instances de classe $2_{\text{diabétiques}}$)
6. Fin

Parallèlement, le recours à la stratégie RESAMPLE modifiée est parvenu, d'une part, à accroître la taille de la classe minoritaire (diabétiques) en la faisant passer de 268 à 384 et, d'autre part, à réduire la taille de la classe majoritaire (non-diabétiques) en la ramenant de 500 à 384, en vue de disposer d'un ensemble de données équilibré de sorte que les deux classes comportent exactement le même nombre d'instances, dans le but de minimiser les biais et de renforcer ainsi la précision de la classification/du diagnostic établi par le classifieur. La figure 4.9 illustre graphiquement l'application de la technique *RESAMPLE* par rapport à la méthode améliorée *IRESAMPLE+*.

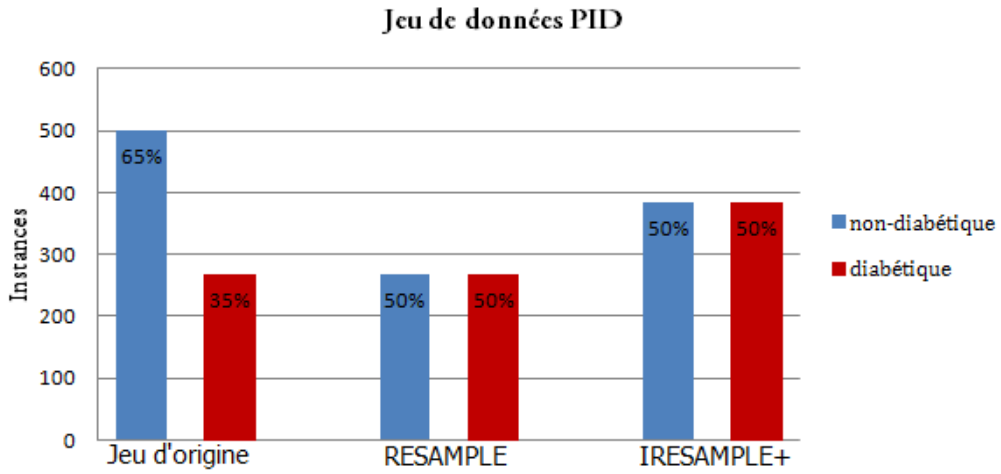

Figure 4.9: Mise en pratique les deux méthodes RESAMPLE et IRESAMPLE+ pour la répartition des classes

3.3.3 Analyse de la performance relative aux modèles

Le présent travail vise à analyser séparément la performance propre à chaque classifieur mis en œuvre, que ce soit avant ou après l'application des techniques de ré-échantillonnage de données (*SMOTE* et *IRESAMPLE+* proposé) et ce, sur la base des critères de performance susmentionnés: la précision (Acc), la sensibilité (Sen), la spécificité (Spe), la valeur prédictive positive (VPP), la valeur prédictive négative (VPN), F-score ou $F-mesure$, le coefficient de corrélation Matthews (CCM) et l'$AUROC$. Les tableaux figurant ci-dessous (tableau 4.3 et tableau 4.4) résument les résultats fournis par les cinq modèles.

Tableau 4.3: Les résultats obtenus au moyen des classifieurs de base de manière indépendante avant de procéder au ré-échantillonnage de l'ensemble PID

Méthode de classification	**Mesures de performance (%)**							
	Acc	Sen	Spe	VPP	VPN	F-score	CCM	AUROC
PMC	75.39	60.80	83.20	66.00	79.80	63.29	44.92	79.30
K-ppv	72.66	55.20	82.00	62.20	77.40	58.49	38.37	74.20
SVM	77.47	55.60	89.20	73.40	78.90	63.27	48.42	72.40
RF	74.87	59.00	83.40	65.60	79.10	62.10	43.51	81.50
NB	76.30	61.20	84.40	67.80	80.20	64.30	46.78	81.90

Tableau 4.4: Les résultats achevés par le biais des classifieurs de base de manière séparée suite au ré-échantillonnage de l'ensemble PID

Filtre	**Méthode de classification**	**Mesures de performance (%)**							
		Acc	Sen	Spe	VPP	VPN	F-score	CCM	AUROC
SMOTE	**PMC**	77.40	79.80	75.00	76.10	78.80	77.90	54.86	83.30
	K-ppv	79.20	84.60	73.80	76.40	82.70	80.30	58.74	83.00
	SVM	79.60	83.40	75.80	77.50	82.00	80.30	59.37	79.60
	RF	80.80	83.80	77.80	79.10	82.80	81.40	61.71	87.80
	NB	76.50	80.80	72.20	74.40	79.00	77.50	53.20	84.60
IRESAMPLE+	**PMC**	98.31	99.00	97.70	97.70	98.90	98.30	96.62	97.30
	K-ppv	98.83	97.90	98.70	98.70	98.00	97.80	97.67	98.90
	SVM	99.37	99.00	99.70	99.70	99.00	99.60	99.22	99.60
	RF	98.96	98.40	98.50	98.50	98.50	98.00	98.22	98.90
	NB	95.44	97.70	93.20	93.50	97.50	95.50	90.98	98.30

Sur la base des résultats présentés aux tableaux 4.3 et 4.4, nous constatons clairement en effet le fait que, après application des techniques de ré-échantillonnage, les performances des classifieurs demeurent toujours supérieures à celles observées dans le cadre des données d'origine. En outre, le filtre *IRESAMPLE+* proposé fournit pour sa part des résultats plus performants que ceux du filtre *SMOTE*; de ce fait, les performances des classifieurs sont considérablement optimisées. La figure 4.10 montre de manière encore plus visible cette analyse comparative utilisant uniquement les valeurs $AUROC$.

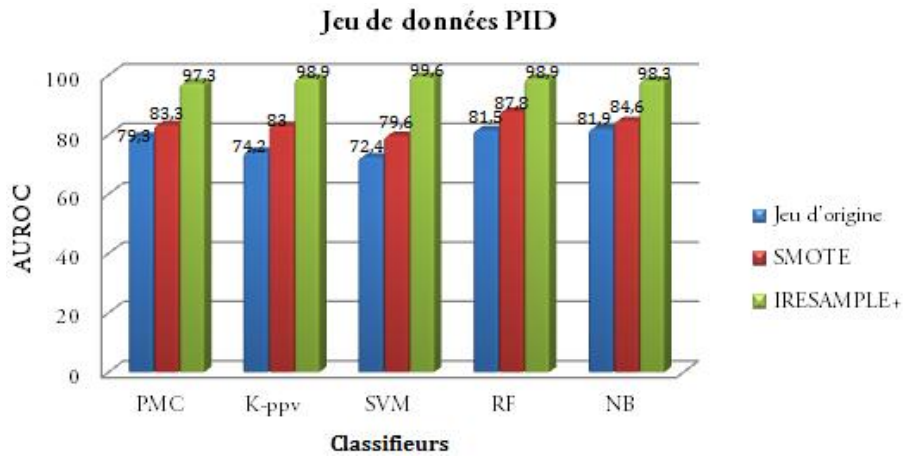

Figure 4.10: Impact du recours aux filtres SMOTE et IRESAMPLE+ en matière de performance des classifieurs

Dans le but notamment d'améliorer la sensibilité et la spécificité dans le cadre de l'apprentissage multimodal proposé ainsi que la performance globale en termes de diagnostic, d'une part, et en vue de parvenir à une décision définitive concernant cette étude, d'autre part, trois approches de fusion sont mises en œuvre: le vote à la majorité (*VM*), le vote pondéré meilleur-pire (*VPMP*) et le méta-classifieur SVM (ou *agrégateur stacking*). Les résultats ainsi obtenus en utilisant les trois techniques *SMOTE*, *RESAMPLE* et *IRESAMPLE+* sont indiqués au tableau 4.5.

Tableau 4.5: Résultats obtenus à travers les trois paradigmes d'agrégation utilisés pour le jeu de données du diabète (PID)

Filtre	**Méthode d'agrégation**	**Mesures de performance (%)**							
		Acc	Sen	Spe	VPP	VPN	F-score	CCM	AUROC
SMOTE	**VM**	81.50	85.40	77.60	79.20	84.20	82.20	63.19	81.50
	VPMP	**82.41**	**84.70**	**79.80**	**80.40**	**84.10**	**82.10**	**65.43**	**82.30**
	SVM méta-classifieur	80.50	84.60	78.40	79.30	81.80	80.90	61.00	80.50
RESAMPLE	**VM**	76.68	75.00	78.40	77.60	75.80	76.70	53.39	76.70
	VPMP	**78.45**	**77.60**	**80.30**	**79.40**	**77.20**	**78.50**	**55.92**	**78.00**
	SVM méta-classifieur	72.02	69.00	75.00	73.40	70.80	72.00	44.11	72.00
IRESAMPLE+	**VM**	98.83	99.70	97.90	98.00	99.70	98.80	97.67	98.80
	VPMP	99.72	100	99.50	99.50	100	99.70	99.46	99.90
	SVM méta-classifieur	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**

En se basant sur les performances décrites ci-dessus (tableau 4.4 et tableau 4.5), nous observons que le concept de l'apprentissage d'ensemble utilisant des techniques de ré-échantillonnage (*SMOTE*, *RESAMPLE* et *IRESAMPLE+*) est nettement préférable par rapport au recours séparé aux classifieurs.

En outre, il convient de souligner que la méthode *VPMP* a généré une classification plus précise en utilisant le filtre *SMOTE & RESAMPLE*, alors que le paradigme d'agrégation de *méta-classifieur SVM* basé sur la stratégie *IRESAMPLE+* est bien plus performant et plus robuste que les autres solutions étudiées dans le cadre de ce travail. Cette approche a également montré une plus grande fiabilité par rapport aux résultats actuels utilisant le même ensemble de données (PID). La figure 4.11 représente les courbes ROC correspondant aux trois méthodes d'agrégation employées -VM, VPMP et le méta-classifieur SVM (*Stacking*)- au moyen du filtre *IRESAMPLE+* suggéré.

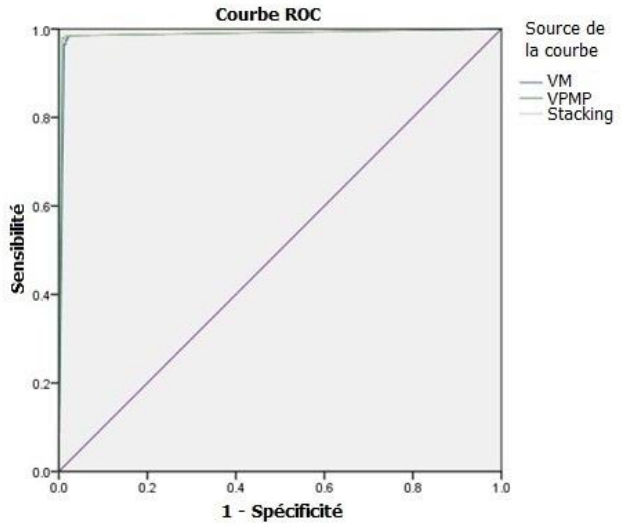

Figure 4.11: Courbes ROC correspondant aux trois schémas d'agrégation utilisés

Il est à noter cependant que tous les points de courbes relatifs à ces trois différentes approches sont positionnés sur la moitié supérieure de la zone ROC, résultant en une meilleure courbe ROC, et en particulier au niveau du modèle de méta-classification SVM.

Les diagrammes présentés par la suite prouvent que l'approche de méta-classification permet de réduire le taux d'erreur de manière notable par rapport à l'apprentissage indépendant des classifieurs. De plus, cette approche présente des performances optimales au regard des critères d'évaluation suivants: précision (*Acc*), sensibilité (*Sen*) et spécificité (*Spe*), qui sont respectivement illustrées à la figure 4.12, figure 4.13 et figure 4.14.

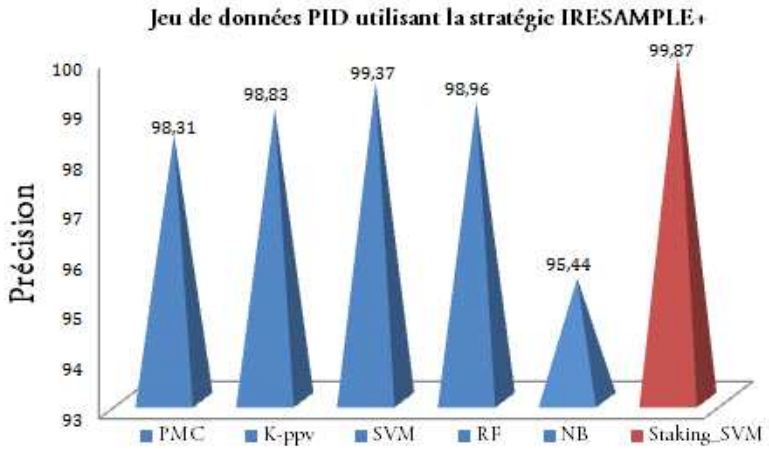

Figure 4.12: Une comparaison entre le taux de précision obtenu au moyen des classifieurs de base et celui de la méthode de méta-classification en utilisant la technique IRESAMPLE+

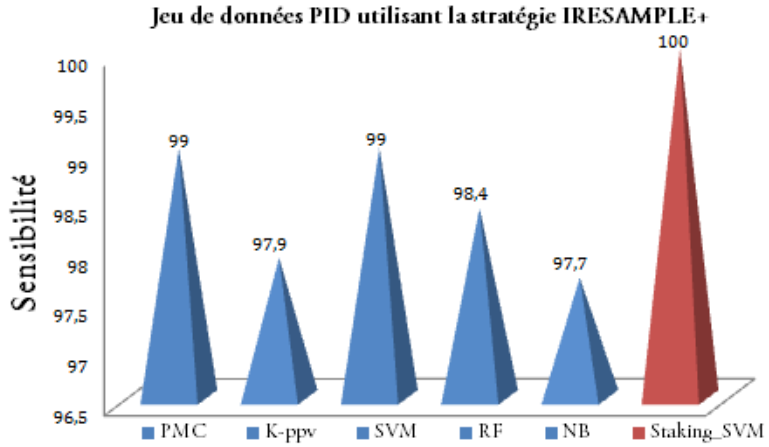

Figure 4.13: Une comparaison entre le taux de sensibilité obtenu au moyen des classifieurs de base par rapport au paradigme de Stacking_SVM utilisant la méthode IRESAMPLE+

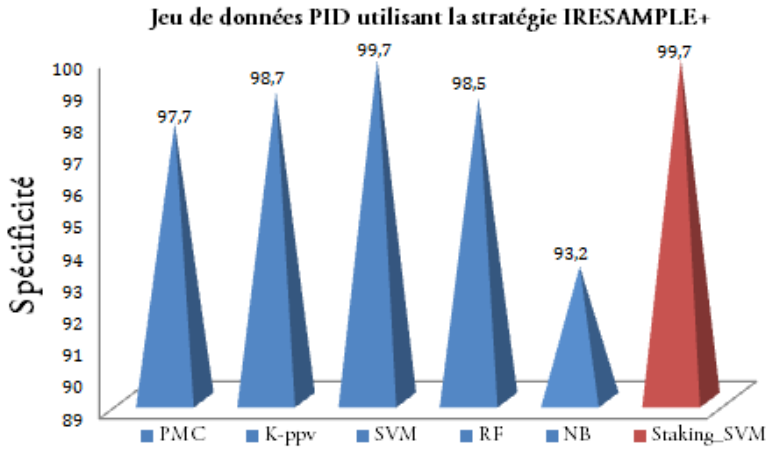

Figure 4.14: Comparaison du taux de spécificité entre les classifieurs de base et l'approche Stacking_SVM à travers le filtre IRESAMPLE+

Il convient également de souligner que les classifieurs individuels "m" (tels que *PMC*, *K-ppv*, *SVM*, *RF* et *NB*) sont désignés comme les apprenants de premier niveau (*classifieurs de base*) tandis que le *méta-classifieur* est identifié comme étant l'apprenant de deuxième niveau.

Dans le cas de chacun de ces classifieurs de premier niveau, des sorties/prédictions sont produites sur la base de l'ensemble de données original "p" relatif à l'apprentissage, tout en créant un nouvel ensemble de données "f" destiné à former le méta-classifieur de deuxième niveau.

Par ailleurs, les prédictions établies par les classifieurs de premier niveau sont transmises au méta-classifieur ou à l'agrégateur stacking sous forme de caractéristiques d'entrée, portant les mêmes étiquettes de classes que celles de l'ensemble de données d'origine. En conséquence, les méta-caractéristiques « *meta-features* » qui en résultent possèdent une taille d'apprentissage équivalente à "$m \cdot f$".

Dans une étude expérimentale traitant des performances du méta-classifieur SVM selon diverses méthodes de noyau utilisées, notamment le noyau *polynomial*, le noyau *RBF* (Radial Basis Function) et le noyau *PUK* en utilisant la méthode de validation croisée $k\ blocs$ au moyen de plusieurs nombres de K (5 et 10). Les résultats de cette étude sont présentés par le tableau 4.6.

Tableau 4.6: Les résultats obtenus par diverses optimisations du noyau dans le cas de deux protocoles $\boldsymbol{K-blocs}$ (**5** et **10**)

k	**Noyau utilisé**	**Mesures de performance (%)**							
		Acc	Sen	Spe	VPP	VPN	F-score	CCM	AUROC
5	Polynomial	99.73	100	99.50	99.50	100	99.70	99.46	99.70
	RBF	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**
	PUK	99.73	99.50	100	100	99.50	99.70	99.46	99.70
10	Polynomial	99.61	99.50	99.70	99.70	99.50	99.60	98.42	99.60
	RBF	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**
	PUK	99.61	99.50	99.70	99.70	99.50	99.60	98.42	99.60

D'après la comparaison des résultats obtenus qui figurent dans le tableau 4.6, il apparaît alors que le *méta-classifieur SVM* au moyen du noyau RBF (*Radial Basis Function*) est le plus performant et ce, indépendamment de la valeur du nombre de K (5 fois/10 fois) utilisée dans le cadre de l'approche de validation croisée, aboutissant ainsi à une précision (Acc) de 99.87% avec une sensibilité (Sen) maximale de 100%, une spécificité (Spe) optimale de 99.70%, une valeur prédictive positive (VPP) de 100%, une valeur prédictive négative (VPN) de 99.50%, un F-score égal à 99.90%, un coefficient de corrélation Matthews (CCM) équivalent à 99.74%, et un AUROC au taux de 99.90%, ce qui permet ainsi de renforcer la performance du diagnostic.

Les principaux paramètres à adapter conformément au méta-classifieur SVM sont spécifiés comme suit: Noyau RBF avec une valeur $gamma$ de 0.5 et le paramètre C égal à 3. En outre, la figure 4.15 propose d'autres mesures importantes permettant d'évaluer la performance du système suggéré que fournit également l'outil Weka, soit la statistique Kappa (représentant la fiabilité de la classification), l'erreur moyenne absolue (EMA/MAE), l'erreur quadratique moyenne (EQM/RMSE), l'erreur relative absolue (ERA/RAE) ainsi que l'erreur quadratique relative (EQR/RRSE), qui sont respectivement définies par les équations (4.5), (4.6), (4.7), (4.8) et (4.9) suivantes:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.5}$$

Où, P_o indique la conformité relative observée entre les évaluateurs de la classification, tandis que P_e est la probabilité hypothétique de la conformité au hasard.

$$EMA = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n} \tag{4.6}$$

$$EQM = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \tag{4.7}$$

$$ERA = \frac{\sum_{i=1}^{n} |P_i - O_i|}{\sum_{i=1}^{n} |\bar{O} - O_i|} \tag{4.8}$$

$$EQR = \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\bar{O} - O_i)^2} \tag{4.9}$$

Où, P_i représente la cible prédite, O_i représente la cible observée, et n correspond au nombre d'observations.

```
Correctly Classified Instances         767               99.8698 %
Incorrectly Classified Instances         1                0.1302 %
Kappa statistic                          0.9974
Mean absolute error                      0.0013
Root mean squared error                  0.0361
Relative absolute error                  0.2604 %
Root relative squared error              7.2168 %
Total Number of Instances              768
```

Figure 4.15: Présentation des taux de statistique Kappa, MAE, RMSE, RAE et RRSE selon la méthode proposée

Le tableau 4.7 suivant représente la matrice de confusion relative à cette étude.

Tableau 4.7: Matrice de confusion de l'approche proposée IRESAMPLE+St

	Test positif (P)	**Test négatif (N)**
Patients diabétiques (P)	384	0
Patients non diabétiques (N)	1	383

En effet, comme l'indiquent la figure 4.15 ainsi que le tableau 4.7, 384 instances de diabète sont correctement identifiées comme des cas de diabète conformément à la méthode proposée, tandis que 383 instances saines sont correctement classées dans les cas de non-diabète. Pour récapituler, 767 instances sont correctement étiquetées alors que seulement une ne l'est pas, résultant en un *coefficient kappa de Cohen* égal à 99.74%, avec une erreur *EMA* égale à 0,13%, une erreur *EQM* équivalente à 3,61%, une erreur *ERA* de 0,2604% et une erreur *EQR* relative à 7,2168%, ce qui démontre encore davantage la fiabilité de notre système de classification du diabète.

Par ailleurs, en analysant les résultats expérimentaux menés au cours de cette étude, nous en sommes parvenus aux conclusions suivantes: le recours à la combinaison de classifieurs est plus avantageux que le fait de procéder à l'apprentissage séparé de ces derniers (réduisant ainsi remarquablement le taux d'erreur), de même que le fait de disposer d'un ensemble de caractéristiques multimodales par le biais de la présente approche de méta-classification

permet désormais de prouver concrètement une efficacité optimale en matière de diagnostic de la maladie du diabète.

La présente contribution vise à mettre au point un système CAD (*Computer-Aided Diagnostic*) de diagnostic de la maladie du diabète plus performant que ceux proposés/existant actuellement dans la littérature et ce, par le biais de la combinaison en une architecture méta-classification des caractéristiques produites au moyen de divers classifieurs ainsi que grâce au procédé $IRESAMPLE +$.

Le tableau 4.8 illustre les approches fondamentales qui sont proposées dans la littérature, tout en présentant une évaluation comparative portant sur la performance relative à la méthode suggérée par rapport aux principales approches existantes en matière de diagnostic du diabète figurant dans la littérature. La présentation de ce tableau est organisée en termes de la mesure de précision (Acc).

Tableau 4.8: Comparaison des performances entre la méthode proposée et les divers modèles liés

Référence des travaux	Jeu de données utilisé	Technique de prétraitement	Méthode de Classification	Acc (%)	Sen (%)	Spe (%)	AUROC (%)
Hasan et al. [149]	Pima Indians Diabetes (PID)	Rejet des valeurs aberrantes, Remplissage des valeurs nulles, Normalisation z-score et Sélection de caractéristiques (corrélation)	Méthode d'ensemble AdaBoost+XGBoost (AB & XB) au moyen du vote pondéré soft	-	78.90	93.40	95.00
Varma et al. [160]	PID	Élimination des valeurs manquantes	Arbre de décision flou gaussien basé sur l'indice de Gini modifié	75.80	-	-	-
Bozkurt et al. [161]	PID	-	Réseaux de temporisation distribués par réseaux de neurones (DTDN)	76.00	53.33	88.75	-
Singh N. & Singh P. [162]	PID	-	Stacking+SMO	79.00	78.90	-	73.20
Choubey & Paul [163]	PID	Sélection de caractéristiques (SC) par l'algorithme génétique	Algorithme génétique + PMC	79.13	79.10	-	84.20
Iyer et al. [164]	PID	Transformation et sélection de caractéristiques	Naïf de Bayes (NB)	79.57	-	-	-

Abdillah & Suwarno [151]	PID	-	SVM + Noyau RBF	80.22	82.56	79.12	80.84
Nai-arun & Moungmai [150]	Hôpital régional Sawanpracharak (HRS)	Transformation et sélection de caractéristiques	Forêts aléatoires	85.56	-	-	91.20
Mahabub [148]	PID	Normalisation	Méthode d'ensemble basée sur le Vote (K-ppv, SVM et PMC)	86.00	-	-	-
Ramezani et al. [165]	PID	Imputation et algorithme de réduction de dimension linéaire OT	Système d'inférence floue basé sur un réseau adaptatif logistique (LANFIS)	88.05	92.15	81.63	-
Alghamdi et al. [166]	Hôpitaux Henry Ford FIT-région métropolitaine de Detroit	Discrétisation, SC (MLR+Entropie) et le filtre SMOTE	Méthode d'ensemble basée sur le Vote	89.00	99.70	74.70	92.20
Nnamoko & Korkontzelos [139]	PID	Technique IQR + SMOTE	C4.5	89.50	89.40	-	94.60
Chen & Pan [167]	Hôpital de l'Univ. médicale WenZhou	La suppression des enregistrements n'est pas au format numérique	LogitBoost	89.63	-	-	96.30
Maniruzzaman et al. [147]	PID	Normalisation à travers la technique médiane	CPG + Noyau RBF	91.97	91.79	63.33	-
Nilashi et al. [168]	PID	Carte auto-organisatrice (SOM) + ACP	Réseau de neurones	92.28	-	-	-
Maniruzzaman et al. [152]	Étude nationale d'examen sanitaire et nutritionnel	Sélection de caractéristiques via régression logistique	Forêts aléatoires	94.25	99.57	-	95.00
Nai-Arun & Sittidech [169]	HRS	SC via le Ratio de Gain	Bagging	95.31	-	-	-
Yilmaz et al. [170]	PID	Algorithme K-means modifié	K-means modifié + SVM	96.71	97.31	95.06	-
Sarwar et al. [146]	-	-	Modèle d'ensemble (K-ppv, SVM, RNA et NB) utilisant le vote à la majorité	98.60	-	-	-
IRESAMPLE+St	PID	Mise en équilibre de données grâce à la stratégie IRESAMPLE+ proposée	Apprentissage multimodal basé sur l'agrégateur stacking via le méta-classifieur SVM	99.87	100	99.70	99.90

En observant le tableau 4.8 à propos des performances relatives aux études connexes figurant dans la littérature, il apparaît en effet que chaque système opère différemment en termes de technique de prétraitement de données ainsi que de méthode de classification adoptée aux fins du diagnostic du diabète au moyen du même ensemble de données, c'est-à-dire le jeu *PID*.

Ainsi, Varma et al. [160] ont mis en place une technique de prétraitement consistant à faire disparaître les valeurs manquantes de l'ensemble de données original, tout en recourant au paradigme de l'arbre de décision flou gaussien basé sur l'indice de Gini modifié à titre d'approche de classification. Bozkurt et al. [161] ont analysé différents classifieurs à base de réseaux de neurones, tels que les réseaux distribués à délai de temporisation, les réseaux d'anticipation, Apprentissage de la quantification vectorielle, les réseaux en cascade, les réseaux de neurones probabilistes ou encore les réseaux à délai de temporisation à partir de l'ensemble des données d'origine. De même, Singh N. & Singh P. [162] ont choisi d'utiliser l'ensemble de données original, à savoir sans aucun prétraitement, tout en adoptant l'approche de généralisation du stacking au moyen de différents noyaux du classifieur SVM du type linéaire, polynomial, RBF et sigmoïde par le biais du méta-apprenant SMO. Choubey & Paul [163] ont opté vers la méthode de sélection des caractéristiques pour le prétraitement du jeu de données, faisant appel à l'algorithme génétique et au classificateur MLP dans le cadre de la phase de diagnostic. Iyer et al. [164] se sont basés sur deux méthodes de prétraitement de données comprenant la transformation ainsi que la sélection de caractéristiques de manière à appliquer le classifieur Naïf de Bayes à cet ensemble résultant. Ramezani et al. [165] ont présenté un système d'inférence floue basé sur le réseau logistique adaptatif (LANFIS) qui utilise la méthode d'imputation ainsi que l'algorithme de réduction de dimension linéaire OT relativement au prétraitement de données. Nilashi et al. [168] ont fait usage la carte auto-organisatrice (SOM) plus la technique ACP en tant que moyen de prétraitement de données ainsi que du paradigme réseau de neurones pour la classification. Yilmaz et al. [170] ont proposé un algorithme de K-means modifié en matière de phase de prétraitement afin de l'utiliser pour la classification au moyen du modèle SVM.

Il convient de mentionner également que la majorité des études ne prennent pas en considération la notion de données déséquilibrées (particulièrement au niveau de données) dans le contexte de la classification du diabète, ce qui constitue pourtant une problématique essentielle au regard du domaine de l'apprentissage automatique, aboutissant ainsi à des résultats de classification erronés. Néanmoins, il existe relativement peu de travaux portant

sur ce concept de déséquilibre de données [139, 166] en utilisant la stratégie SMOTE. Pareillement, la majorité des approches citées par la littérature comme étant des méthodes de diagnostic du diabète sont bien souvent basées sur une approche standard qui dépend du point de vue d'un classificateur unique ou encore sur une approche d'ensemble par vote (agrégation de type classe) fournissant ainsi un nombre réduit d'informations (disponibles) lors de la fusion.

L'approche proposée *IRESAMPLE+St* par la présente étude aborde le concept de données non équilibrées aux deux niveaux, autrement dit au niveau des données par le biais de la proposition d'une méthode améliorée de RESAMPLE désignée par le terme *IRESAMPLE+*, ainsi qu'au niveau des algorithmes au moyen du recours à une approche d'ensemble multimodale basée sur un agrégateur stacking (incluant la technique d'apprentissage croisé).

En comparant les résultats présentés dans le tableau 8 et par rapport aux études mentionnées au cours de ce chapitre, la méthode *IRESAMPLE+St* surpasse largement celles de l'état de l'art en obtenant les résultats les plus optimaux en termes de précision (Acc), de sensibilité (Sen), spécificité (Spe), AUROC, et par le kappa de Cohen. Cela confirme également que la méthode de méta-classification suggérée *IRESAMPLE+St* offre un diagnostic précoce meilleur et plus précis de la maladie du diabète avec $Acc = 99.87\%$, $Sen = 100\%$, $Spe = 99.70\%$, $AUROC = 99.90\%$ et $kappa\ de\ Cohen = 99.74\%$.

4 Conclusion

Aujourd'hui, l'apprentissage automatique s'est considérablement développé dans le domaine du diagnostic médical, notamment en ce qui concerne le diagnostic du diabète, et à ce titre, grâce à l'intégration du concept d'apprentissage déséquilibré, qui peut générer des résultats de classification erronés. Ce concept est traité sous deux angles, c'est-à-dire au niveau des données via la modification/équilibrage du jeu de données d'apprentissage ainsi qu'au niveau des algorithmes.

Dans ce chapitre, nous avons adopté une approche hybride vers l'apprentissage déséquilibré en proposant une méthode de méta-classification multimodale améliorée appelée *IRESAMPLE+St* sur la base du paradigme Stacking pour distinguer les patients normaux et diabétiques. Dans le même objectif de cette étude, une technique modifiée basée sur RESAMPLE appelée *IRESAMPLE+* est intégrée comme étape de ré-échantillonnage préliminaire en vue de surmonter le problème de déséquilibre de données.

En réalité, le recours à la technique de ré-échantillonnage *IRESAMPLE+* en traitant la notion de déséquilibre des données par le biais d'une approche ensembliste « *méta-classification multimodale*» constitue une solution plus performante avec des résultats impressionnants en les comparant aux principales études connexes et ce, dans le but de fournir un système d'aide à la décision clinique plus robuste qui permet aux diabétologues de diagnostiquer rapidement les patients diabétiques à un stade précoce et de soutenir leur décisions thérapeutiques.

CONCLUSION GÉNÉRALE ET PERCEPTIVES

Conclusion Générale et Perceptives

La classification est un problème bien connu dans les communautés d'apprentissage automatique. La précision de la classification est principalement étudiée à partir d'une seule source d'information. Cependant, prendre une décision en s'appuyant sur une source unique est susceptible de compromettre le processus de prise de décision. Le recours à des données multimodales ainsi qu'à l'apprentissage profond constitue des pistes envisageables visant ainsi à renforcer la précision des systèmes d'aide à la décision. En effet, l'apprentissage profond dans le contexte de l'apprentissage multimodal constitue un domaine très prometteur comportant de nombreuses applications concrètes, notamment le diagnostic médical et la classification de textes. Cette approche multimodale se fonde principalement sur la fusion des informations issues de diverses modalités dans le but de déterminer la représentation combinée la plus appropriée de celles-ci pour permettre une prise de décision plus efficace.

La fusion multimodale comprend essentiellement deux catégories importantes, à savoir une fusion au niveau de caractéristiques ainsi qu'une fusion au niveau de classifieurs. Dans le cadre du processus de la fusion des caractéristiques, nous disposons de caractéristiques en entrée et en sortie. Le but étant de créer ou de renforcer un nouvel ensemble de caractéristiques provenant de diverses sources. Ainsi, la manière la plus aisée en termes de cette fusion consiste à concaténer ces caractéristiques en un vecteur unique. Donc, la phase de classification s'effectue par rapport au nouvel échantillon de caractéristiques résultant de la fusion. Néanmoins, une telle approche requiert une certaine synchronisation entre ces modalités. Par ailleurs, dans le cas de fusion des classifieurs, il est considéré en effet que les modalités étant indépendantes, il convient donc dans un premier temps de mettre en œuvre des classifieurs séparés pour chaque modalité, puis de procéder à la fusion par le biais d'un module d'agrégation permettant de fusionner les différents sorties. Ainsi, la solution la plus pratique du point de vue de cette fusion est de recourir au vote afin de parvenir à la décision finale. En revanche, ce module a seulement recours aux décisions des systèmes monomodaux.

En outre, la plupart des études portant sur la fusion d'informations reposent essentiellement sur la fusion de classifieurs. Cependant, il est impossible d'exploiter entièrement la corrélation entre les diverses sources/ modalités de données, étant donné que chaque classifieur est local et autonome par rapport aux autres. Parallèlement, la fusion au niveau de caractéristiques se

révèle être plus avantageuse que la fusion au niveau de classification en présence de modalités étroitement liées entre elles.

Il convient toutefois à présent de déterminer la manière de modéliser au mieux les données multimodales ou la mise en pratique de la notion de multimodalité au moyen de l'apprentissage automatique, y compris l'apprentissage profond, et ce, en vue d'améliorer les performances au niveau de divers systèmes d'aide à la décision.

Dans le but de surmonter toutes ces contraintes de manière à présenter une solution optimisée, la présente thèse se concentre sur l'analyse de l'impact ainsi que la mise en œuvre d'une approche multimodale appliquée dans le cadre de nombreuses sources d'informations d'une part (les données textuelles et celles issues de l'imagerie médicale) et d'autre part selon la nature des bases traitées (équilibrées ou non).

Dans un premier lieu, nous avons recours à une stratégie de classification multimodale reposant sur les réseaux de neurones convolutionnels (*CNNs*) et le classifieur *SVM*. D'une part, dans le contexte textuel, nous avons adopté une démarche multimodale du niveau décision en matière de vérification des auteurs par le biais de la méthode de plongement lexical «*Word Embedding*». Dans un premier stade, nous avons réalisé une étude comparative portant en particulier sur le choix (en termes de performance) entre la combinaison de classifieurs par rapport à leur apprentissage séparé ou l'apprentissage monomodal.

De ce fait, les expérimentations menées à bien au cours de cette approche ont abouti à proposer la mise en place d'une nouvelle approche multi-classifieurs qui, basée sur le principe de fusion tardive, fait appel à trois architectures différentes - *CNN*, *RCNN* et *SVM* - pour la tâche de classification des textes d'auteurs selon leur style d'écriture en utilisant le *Word2vec*. De même, cette approche consiste également à faire une analyse par rapport au comportement du système multimodal mis au point conformément au modèle *Word2vec* adopté, soit *Skip-Gram* ou *CBOW*.

Simultanément, plusieurs expérimentations portant sur l'évaluation du système multi-classifieurs proposé ont été effectuées au moyen de différentes règles en matière d'agrégation de type classe, à savoir le vote à la majorité (*VM*), le vote à la majorité pondérée (*VMP*) et le vote pondéré meilleur-pire (*VPMP*), et ce, en vue de parvenir à une décision finale optimale. De fait, il apparaît après analyse des résultats ainsi obtenus, que la combinaison de classifieurs selon le modèle *Skip-Gram* est plus appropriée que le modèle *CBOW* qui le dépasse largement

et ce dans tous les cas. En outre, la technique VPMP se révèle être une solution plus performante par rapport aux règles VM et VMP, avec des résultats très encourageants.

Par ailleurs, nous suggérons, dans le cadre de l'imagerie médicale, une nouvelle méthode de classification multimodale impliquant la fusion de classifieurs sur la base à la fois de données multimodales et de caractéristiques diversifiées pour le diagnostic du glaucome. A ce propos, ce travail constitue le premier qui propose deux niveaux de multimodalité (*le niveau modalités et/ou caractéristiques et le niveau décision*) en introduisant une nouvelle approche en matière de fusion multimodale, connue sous le nom *d'approche hybride*, permettant de mettre à profit la coopération de deux concepts de fusion précoce et tardive; Ainsi, d'une part, nous mettons en concatène les caractéristiques propres aux différentes modalités préalablement au processus d'apprentissage et, d'autre part, nous appliquons des classifieurs séparés/distincts à chaque combinaison de modalités de manière à assurer réellement cette notion de multimodalité. Par la suite, les diverses sorties sont fusionnées au moyen de trois méthodes de vote communément utilisées -*VM*, *VMP* et *VPMP*- permettant ainsi de déterminer la décision finale relative au système multimodal proposé. À cet effet, la règle VPMP fournissait une performance optimale au terme de plusieurs évaluations. En outre, la combinaison de nombreuses caractéristiques multimodales s'est avérée spécialement performante dans le diagnostic du glaucome.

Toutefois, lors de ce travail, nous avons procédé à une évaluation comparative par rapport aux performances du classifieur SVM à travers la conception de différentes combinaisons possibles en matière de familles de caractéristiques. Les résultats ainsi achevés indiquent une supériorité considérable de la combinaison des trois techniques traditionnelles d'extraction de caractéristiques (*GLCM, Moments centraux et Moments Hu*) associée aux caractéristiques obtenues par voie automatique grâce au réseau *CNN* vis-à-vis de la combinaison séparée de ces trois dernières caractéristiques avec celles fournies par *CNN*.

En revanche, une étude expérimentale est menée également au sujet de l'impact du nombre d'époques sur la performance des réseaux *CNNs* selon la fonction d'activation utilisée « *ReLU, Identité, Sigmoid ou* Tanh ». Globalement, nous avons remarqué que l'augmentation en nombre d'époques se traduit en une diminution du taux d'erreur au niveau des deux réseaux CNNs ($CNN1_{RVB}$ et $CNN2_{Binaire}$), et ce, indépendamment de la fonction d'activation choisie. D'après les résultats obtenus, la performance maximale des réseaux *CNNs* est atteinte dès que le nombre d'époques équivaut à soixante, alors que la précision est pratiquement stable après

cinquante-cinq époques. En outre, nous constatons également que le modèle $CNN1_{RVB}$ offre une précision optimale au moyen de la fonction *ReLU*, alors que le modèle $CNN2_{Binaire}$, via la fonction *Identité*, offre des performances supérieures à toutes les autres solutions envisagées, tout en présentant des résultats plus favorables.

À l'issue de l'analyse des résultats obtenus dans le cadre de cette étude, il ressort que l'apport de l'apprentissage multimodal utilisant le réseau CNN et le classifieur SVM dans l'amélioration de la performance de l'approche proposée se révèle être extrêmement remarquable. En effet, la phase d'agrégation du niveau de décision au moyen de la règle *VPMP* constitue celle qui présente par excellence les meilleurs aboutissements, et ce de manière particulièrement satisfaisante en termes de précision, de sensibilité et de spécificité.

Dans un second lieu, nous avons opté en faveur pour une démarche hybride, dans le cadre du déséquilibre de données, de manière à mettre au point une méthode de méta-classification multimodale améliorée, à savoir celle de l'*IRESAMPLE+St*, et ce, aux fins du diagnostic de la maladie du diabète. En outre, nous avons traité ce concept d'apprentissage déséquilibré selon deux aspects distincts, c'est-à-dire au niveau des données en modifiant/équilibrant l'ensemble de données relatives à l'apprentissage, de même qu'au niveau des algorithmes.

Pour ce qui est du premier objectif, et en particulier en vue de surmonter le problème des données non-équilibrées, cette étude a porté sur une stratégie modifiée basée sur la méthode RESAMPLE, désignée sous le terme d'*IRESAMPLE+*, ainsi que sur la méthode SMOTE; ces deux méthodes sont appliquées à titre d'étape préliminaire dans le cadre du processus de ré-échantillonnage de manière à offrir une meilleure représentativité de ces données durant la phase de classification. Par ailleurs, au sujet du second volet, nous avons adopté une approche ensembliste, notamment le paradigme *Stacking* dans la mesure où il est possible de combiner différentes sorties/caractéristiques propres à plusieurs modèles de classification à travers un méta-classifieur de haut niveau (SVM), de manière à diminuer la variance, le biais, et/ou à renforcer les prédictions.

Au cours de ce travail, nous avons procédé à diverses expériences relatives à la constitution du pool de classifieurs de base et, plus précisément, au nombre de classifieurs choisis, à travers l'adoption de cinq modèles distincts (*PMC, K-ppv, SVM, RF* et *NB*). Cette démarche correspond à la génération d'un ensemble de classifieurs hétérogènes/complémentaires (*multimodaux*), dont la combinaison (*méta-apprentissage*) permettant ainsi de parvenir à une solution optimale.

Parallèlement, le recours à la stratégie IRESAMPLE+ en regard de la technique classique RESAMPLE a en effet permis de disposer d'un jeu de données plus équilibré, de sorte que les deux classes possèdent un nombre identique d'instances, de manière à minimiser le biais et ainsi à renforcer la précision de la classification. En règle générale, les résultats expérimentaux fournis dans le cadre de ce travail ont clairement montré que, suite à la mise en pratique des techniques de ré-échantillonnage, les performances des classifieurs de base ainsi que leur agrégation sont toujours supérieures à celles obtenues avec les données originales. En outre, la stratégie IRESAMPLE+ a apporté des résultats plus performants que ceux du filtre SMOTE. Il convient également de noter que l'apprentissage multimodal ensembliste reste largement prépondérant par rapport à celui monomodal.

De plus, l'étude comparative établie par rapport à d'autres méthodes de fusion de type classe (VM et VPMP) a permis de prouver une fois de plus que la stratégie d'agrégateur stacking ou de méta-classifieur SVM est plus performante. Ainsi, d'après l'étude empirique à propos de diverses optimisations du noyau au niveau de méta-classifieur SVM via deux protocoles de validation croisée $K-blocs$ (5 et 10), il est alors apparu que le *méta-classifieur SVM* au moyen du noyau *RBF* est le plus approprié, peu importe la valeur du nombre de K (5 fois/10 fois), conduisant ainsi à des résultats optimaux. Enfin, l'approche *IRESAMPLE+St* s'est révélée être une solution optimale par rapport aux autres méthodes similaires proposées dans le contexte du déséquilibre des données.

Les résultats expérimentaux menés dans le cadre de cette thèse confirment que la contribution du concept multimodal (selon la nature des bases traitées équilibrées ou non) aux divers systèmes actuels d'aide à la décision basés sur l'apprentissage profond demeure incontournable/primordiale, ce qui permet également de fournir des résultats impressionnants en les comparant aux principales études connexes qui nous permettent de poursuivre dans cette voie de recherche.

Perspectives

Les résultats obtenus par nos approches sont très impressionnants, ils nous encouragent à poursuive cette voix de la recherche. De nombreuses perspectives peuvent être envisagées à la suite de cette thèse, telles que:

- Face au problème majeur des limites de données et du choix adéquat en matière d'hyper-paramètres dans le cadre de l'apprentissage des réseaux de neurones profonds, il s'avère délicat du point de vue de l'apprentissage optimal de ces réseaux. De ce fait, il serait avantageux d'introduire l'idée de l'augmentation des données ainsi que de mettre en place également d'autres modèles prédéfinis et préformés faisant appel à l'apprentissage par transfert.

- Connaissant clairement l'avantage lié à la sélection de caractéristiques dans le cadre d'un problème de classification, nous pouvons étudier la possibilité de combiner diverses techniques de sélection de caractéristiques, y compris SBC (*Selective Bayesian classifier*) et mRMR (*Maximum Relevance Minimum Redundancy*), avec des réseaux CNNs dans le but de ne conserver que les plus représentatives / utiles pour la classification. La construction de cartes de caractéristiques doit être améliorée pour étudier les caractéristiques de niveau plus profond. De cette manière, le réseau peut collecter des caractéristiques plus détaillées en vue de renforcer la capacité de discrimination.

- l'enrichissement de l'espace des caractéristiques serait très intéressant. L'utilisation d'autres caractéristiques de texture et de forme associées à de nouvelles modalités pour représenter la zone d'intérêt ainsi que des caractéristiques de couleurs afin d'extraire plus d'informations des images couleurs telles que les images du fond d'œil.

- Afin d'exploiter durablement notre système, il serait intéressant que chaque image/instance bien classée et approuvée par le médecin soit intégrée/ajoutée à la base d'apprentissage d'origine. Dans ce cas, le système proposé pourra reconnaître de nouveaux cas (cas inconnus).

- Le recours à un algorithme de sélection dynamique de classifieurs en vue d'équilibrer le compromis entre la précision et la diversité du pool, ainsi que pour étudier l'impact de l'utilisation de critères de sélection dans le choix de l'ensemble initial des classifieurs.

- Mettre en place un système adapté aux autres problématiques relatives à l'apprentissage déséquilibré dans le cadre de la classification multi-label en abordant principalement ce défi au niveau des algorithmes.

PRODUCTIONS SCIENTIFIQUES

Productions Scientifiques

Publications dans des revues internationales

[1] N.E. Benzebouchi, N. Azizi, A.S. Ashour, N. Dey, R.S. Sherratt, "Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis", Journal of Experimental & Theoretical Artificial Intelligence, Taylor & Francis, 31(6), p. 841–874, 2019

[2] N.E. Benzebouchi, N. Azizi, S.E. Bouziane, "Glaucoma Diagnosis Using Cooperative Convolutional Neural Networks", International Journal of Advances in Electronics and Computer Science, vol.5(1), pp. 31-36, 2018

[3] R. Touahri, N. Azizi, N.E. Hammami, M. Aldwairi, N.E. Benzebouchi, O. Moumene, "Multi source Retinal Fundus Image Classification Using Convolution Neural Networks Fusion and Gabor-Based Texture Representation", International Journal of Computational Vision and Robotics, inderscience, 2020, in press

[4] N.E. Benzebouchi, N. Azizi, D. Schwab, S.B. Belhaouaric, N. Zemmal, "Unbalanced Multimodal Learning for Early Automatic Diagnosis of Diabetes Based on Enhanced Resampling Technique and Stacking Classifier", 2020.

Chapitres d'ouvrages scientifiques

[5] N.E. Benzebouchi, N. Azizi, K. Ayadi, "A Computer-Aided Diagnosis System for Breast Cancer Using Deep Convolutional Neural Networks", In: Behera H., Nayak J., Naik B., Abraham A. (eds) Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, vol 711. Springer, Singapore, 2019

[6] M. Lamari, N. Azizi, N.E. Hammami, S. Cheriguene, A. Boukhamla, N. Dendani, N.E. Benzebouchi, "SMOTE-ENN Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification", Advances in Intelligent Systems and Computing in Advances on Smart and Soft Computing, Springer, vol.1188, 2021

Publications dans des conférences internationales

[7] N.E. Benzebouchi, N. Azizi, S.E. Bouziane, "Glaucoma Diagnosis Using Cooperative Convolutional Neural Networks", International Conference on Control, Automation, Robotics and Vision Engineering (ICCARVE), Rome, Italy, 2017

[8] N.E. Benzebouchi, N. Azizi, M. Aldwairi, N. Farah, "Multi-classifier system for authorship verification task using word embeddings", 2nd International Conference on Natural Language and Speech Processing (ICNLSP), IEEE, Algiers, Algeria, 2018

[9] R. Touahri, N. Azizi, N.E. Benzebouchi, N.E. HAMMAMI, O. Moumene, "A Comparative Study of Convolutional Neural Network and Twin SVM for Automatic Glaucoma Diagnosis", International Conference on Signal, Image, Vision and Their Applications (SIVA), IEEE, Guelma, Algeria, 2018

[10] N.E. Benzebouchi, N. Azizi, N.E. Hammami, D. Schwab, M.C.E. Khelaifia, M. Aldwairi, "Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector", 16th International Multi-Conference on Systems, Signals and Devices (SSD-CSP), IEEE, Istanbul, Turkey, 2019

[11] N. Zemmal, N. Azizi, A. Ziani, N. E. Benzebouchi, M. Aldwairi, "An Enhanced Feature Selection Approach based on Mutual Information for Breast Cancer Diagnosis", 2019 6th International Conference on Image and Signal Processing and Their Applications (ISPA), IEEE, Mostaganem, Algeria, 2019

BIBLIOGRAPHIE

Bibliographie

[1] McCarthy, J., Minsky, M., Rochester, N., Shannon, C. E.: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf, 1955

[2] Mitchell, M. T.: Machine Learning. McGraw-Hill, 1997.

[3] Russell, S. J., & Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, 2009.

[4] Pan, S. J., & Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), p.1345–1359, 2010

[5] Cortes, C., & Vapnik, V.: Support-vector networks. Machine Learning, 20(3), p.273–297, 1995

[6] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks, 61, p. 85–117, 2015

[7] Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65(6), 1958. "Cornell Aeronautical Laboratory"

[8] Benzebouchi, N. E., Azizi, N., Ashour, A. S., Dey, N., & Sherratt, R. S.: Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis. Journal of Experimental & Theoretical Artificial Intelligence, 31(6), p.841-874, 2019

[9] Schuster, M., & Paliwal, K. K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), p.2673–2681, 1997

[10] Hochreiter, S., & Schmidhuber, J.: Long Short-Term Memory. Neural Computation, 9(8), p.1735–1780, 1997

[11] Schuster, M., & Paliwal, K. K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), p.2673–2681, 1997

[12] Benzebouchi, N. E., Azizi, N., & Ayadi, K. (2019). A computer-aided diagnosis system for breast cancer using deep convolutional neural networks. In H. Behera, J. Nayak, B. Naik, & A. Abraham (Eds.), Computational intelligence in data mining. Advances in intelligent systems and computing (Vol. 711, pp. 583–593). Singapore: Springer.

[13] Yang, X., Liu, C., Wang, Z., Yang, J., Min, H.L., Wang, L., & Cheng, K.-T.: Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Medical Image Analysis, 42, pp.212–227, 2017

[14] Oh, S.L., Ng, E.Y.K., Tan, R.S., & Acharya, U.R.: Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in Biology and Medicine, 102, pp.278-287, 2018

[15] Kaya, D.: The mRMR-CNN based Influential Support Decision System Approach to Classify EEG Signals. Measurement, 156, 107602, 2020

[16] Liu, G., & Guo, J.: Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing, 337, pp.325-338, 2019

[17] Nedjah, N., Santos, I. & de Macedo Mourelle, L.: Sentiment analysis using convolutional neural network via word embeddings, Evolutionary Intelligence, 2019. https://doi.org/10.1007/s12065-019-00227-4

[18] Jain, G., Sharma, M., & Agarwal, B.: Spam detection in social media using convolutional and long short term memory neural network. Annals of Mathematics and Artificial Intelligence, 85, pp. 21–44, 2019

[19] Saçak, B.: Media Literacy in a Digital Age: Multimodal Social Semiotics and Reading Media. In Yildiz, M. N., Fazal, M., Ahn, M., Feirsen, R., & Ozdemir, S. (Eds.), Handbook of Research on Media Literacy Research and Applications Across Disciplines, IGI Global, p. 13-26, 2019

[20] Lahat, D., Adali, T., & Jutten, C.: Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proceedings of the IEEE, 103(9), p.1449–1477, 2015

[21] Yilmaz, T., Yazici, A. & Kitsuregawa, M.: RELIEF-MM: effective modality weighting for multimedia information retrieval. Multimedia Systems, 20, p.389–413, 2014

[22] Duin, R. P. W.: The combining classifier: to train or not to train?. Object Recognition Supported by User Interaction for Service Robots, Quebec City, Quebec, Canada, 2, p.765 -770, 2002

[23] Kludas J., Bruno E., Marchand-Maillet S. (2008) Information Fusion in Multimedia Information Retrieval. In: Boujemaa N., Detyniecki M., Nürnberger A. (eds) Adaptive Multimedia Retrieval: Retrieval, User, and Semantics. AMR 2007. Lecture Notes in Computer Science, vol 4918, p.147-159, Springer, Berlin, Heidelberg

[24] YANG, M.-H., & TAO, J.-H.: Data fusion methods in multimodal human computer dialog. Virtual Reality & Intelligent Hardware, 1(1), p.21–38, 2019

[25] Xu, L., Krzyzak, A., & Suen, C. Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics, 22(3), p.418–435, 1992

[26] Bokhari, H.U., & Hasan, F.: Multimodal information retrieval: Challenges and future trends. International Journal of Computer Applications, 74(14), p.9–12, 2013.

[27] Xie, Z., & Guan, L.: Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis. 2012 IEEE International Symposium on Multimedia, Irvine, CA, p. 1-8, 2012, doi: 10.1109/ISM.2012.9.

[28] Faundez-Zanuy M. (2009) Data Fusion at Different Levels. In: Esposito A., Hussain A., Marinaro M., Martone R. (eds) Multimodal Signals: Cognitive and Algorithmic Issues. Lecture Notes in Computer Science, vol 5398, p.94-103, Springer, Berlin, Heidelberg

[29] Ho, T.K., Hull, J.J., & Srihari, S.N. (1992) Combination of Decisions by Multiple Classifiers. In: Baird H.S., Bunke H., Yamamoto K. (eds) Structured Document Image Analysis. Springer, Berlin, Heidelberg

[30] Wu, Y., Chang, E. Y., Chang, K. C.-C., & Smith, J. R.: Optimal multimodal fusion for multimedia data analysis. Proceedings of the 12th Annual ACM, New York, USA, p. 572-579, 2004

[31] Dong, J., Zhuang, D., Huang, Y., & Fu, J.: Advances in Multi-Sensor Data Fusion: Algorithms and Applications. Sensors, 9(10), p.7771–7784, 2009

[32] Anitha, R., Renuka, S., & Abudhahir, A.: Multi sensor data fusion algorithms for target tracking using multiple measurements. 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, India, 2013

[33] Ross A. (2009) Fusion, Feature-Level. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA

[34] Chibelushi, C.C., Mason, J.S.D., & Deravi, F.: Feature-level data fusion for bimodal person recognition. In: Proceedings of the Sixth International Conference on Image Processing and Its Applications, Dublin, Ireland, vol. 1, p. 399–403, 1997

[35] Du, C., Wang, Y., Wang, C., Shi, C., & Xiao, B.: Selective feature connection mechanism: Concatenating multi-layer CNN features with a feature selector. Pattern Recognition Letters, 129, p. 108-114, 2020

[36] Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), p.226–239, 1998

[37] Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), p.21–45, 2006

[38] Kanal, L.: Patterns in pattern recognition. IEEE Transactions on Information Theory, 20(4), p.697-722, 1974

[39] Wolpert, D. H.: The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), p.1341–1390, 1996

[40] Breiman, L.: Bagging predictors. Machine learning, 24(2), p.123–140, 1996

[41] Freund, Y.: Boosting a weak learning algorithm by majority. Information and computation, 121(2), p.256–285, 1995

[42] 42. Breiman, L.: Random forests. Machine learning, 45(1), p.5–32, 2001

[43] Dietterich, T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol.1857. Springer, Berlin, Heidelberg

[44] Vapnik, V. N.: An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), p.988–999, 1999

[45] Bishop, C. M.: Pattern recognition and machine learning, information science and statistics. Springer-Verlag New York, 2006

[46] Nguyen, T. T., Dang, M. T., Baghel, V. A., Luong, A. V., McCall, J., & Liew, A. W.-C.: Evolving interval-based representation for multiple classifier fusion. Knowledge-Based Systems, 201, p.1-20, 2020

[47] Jain, A., Duin, R., & Mao, J.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), p.4-37, 2000

[48] Ruta, D. and Gabrys, B.: An overview of classifier fusion methods. Computing and Information Systems, 7, p.1-10, 2000

[49] Tharwat, A.: Classification Assessment Methods. Applied Computing and Informatics, 2018, In Press

[50] Boser, B. E., Guyon, I. M., & Vapnik, V. N. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational learning theory, ACM, p.144–152, 1992

[51] Bradley, A. P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), p.1145–1159, 1997

[52] Gama, J., & Brazdil, P.: Cascade Generalization. Machine Learning, 41, p.315–343, 2000

[53] Kuncheva, L. I.: A theoretical study on six classifier fusion strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), p.281–286, 2002

[54] Safon, G., Salazar, A., Vergara, L.: Vector score alpha integration for classifier late fusion. Pattern Recognition Letters, 136, p.48-55, 2020

[55] Asker, L., & Maclin, R.: Ensembles as a Sequence of Classifiers. Proceedings of the Fifteenth international joint conference on Artifical intelligence, vol.2, p.860–865, 1997

[56] Kuncheva, L. I.: Clustering-and-selection model for classifier combination. KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, Brighton, UK, vol.1 p.185-188, 2000

[57] Duin, R.P.W., & Tax D.M.J. (2000) Experiments with Classifier Combining Rules. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg

[58] Zouari, H., Heutte, L., Lecourtier, Y., & Alimi, A.: Un panorama des méthodes de combinaison de classifieurs en reconnaissance de formes. In Proc.RFIA, Angers, France, vol.2, p.499-508, 2002

[59] Ho, T. K., Hull, J. J., & Srihari, S. N.: Decision combination in multiple classifier systems. IEEE transactions on pattern analysis and machine intelligence, 16(1), p.66–75, 1994

[60] Lam, L., & Suen, S. Y.: Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 27(5), 1997

[61] Cordella, L.P., Foggia, P., Sansone, C., Tortorella, F., & Vento, M. (1998) Optimizing the error/reject trade-off for a multi-expert system using the Bayesian combining rule. In: Amin, A., Dori, D., Pudil, P., & Freeman, H. (eds) Advances in Pattern Recognition. Lecture Notes in Computer Science, vol.1451. Springer, Berlin, Heidelberg

[62] Huang, Y. S., & Suen, C. Y.: A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(1), p.90 - 94, 1995

[63] Watson, K.B. (2014) Categorical Data Analysis. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht

[64] Grabisch, M.: On equivalence classes of fuzzy connectives-the case of fuzzy integrals. IEEE Transactions on Fuzzy Systems, 3(1), p.96-109, 1995

[65] Benzebouchi, N. E., Azizi, N., Aldwairi, M., & Farah, N.: Multi-classifier system for authorship verification task using word embeddings. 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), IEEE, Algiers, Algeria, p.1-6, 2018

[66] Koppel, M., & Winter, Y.: Determining if two documents are by the same author. JASIST, 65 (1), p.178-187, 2014

[67] Koppel, M., & Schler, J.: Authorship verification as a one-class classification problem. Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, ACM, p.62-69, 2004

[68] Iqbal, F., Khan, L.A., Fung, B.C.M., & Debbabi, M.: E-mail authorship verification for forensic investigation Proceedings of the 2010 ACM Symposium on Applied Computing, New York, USA, ACM, p.1591-1598, 2010

[69] Brocardo, M. L., Traore, I., Saad, S., & Woungang, I.: Authorship verification for short messages using stylometry. Proceedings of the International Conference on Computer, Information and Telecommunication Systems, IEEE, Piraeus-Athens, Greece, p.1-6, 2013

[70] Brocardo, M. L., Traore, I., & Woungang, I. (2015). Authorship verification of e-mail and tweet messages applied for continuous authentication. Journal of Computer and System Sciences, 81(8), p.1429-1440, 2015

[71] Maitra, P., Ghosh, S., & Das, D.: Authorship Verification: An Approach based on Random Forest: Notebook for PAN at CLEF 2015. ArXiv, abs/1607.08885

[72] Mikolov, T., Yih, W.t., & Zweig, G.: Linguistic regularities in continuous space word representations. In NAACL-HLT, p. 746–751, 2013

[73] Zhang, X., Zhao, J., & LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems, p.649–657, 2016, arXiv:1509.01626

[74] Benzebouchi, N.E., Azizi, N., Hammami, N.E., Schwab, D., Khelaifia, M.C.E., & Aldwairi, M.: Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector. 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, IEEE, p. 371-376, 2019

[75] Moreno-Seco, F., Iñesta, J. M., de León, P. J. P., & Micó, L. (2006). Comparison of Classifier Fusion Methods for Classification in Pattern Recognition Tasks. In: DY. Yeung, J. T. Kwok, A. Fred, F. Roli , & D. de Ridder (Eds.) Structural, Syntactic, and

Statistical Pattern Recognition. Lecture Notes in Computer Science (Vol. 4109, p.705–713), Springer, Berlin, Heidelberg.

[76] Onan, A., Korukoğlu, S., & Bulut, H.: A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications, 62, p.1-16, 2016

[77] Zhang, Y., Wallace, B.C.: A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, 2016. arXiv:151003820

[78] Quigley, H. A., & Broman, A. T.: The number of people with glaucoma worldwide in 2010 and 2020. Br. J. Ophthalmol, 90(3), p.262–267, 2006

[79] Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph, 31, p.198–211, 2007

[80] Zemmal, N., Azizi, N., Dey, N., & Sellami, M.: Adaptative Semi Supervised Support Vector Machine Semi Supervised Learning with Features Cooperation for Breast Cancer Classification. Journal of Medical Imaging and Health Informatics, 6(1), p.53-62, 2016

[81] Zemmal, N., Azizi, N., Sellami, M., Zenakhra, D., Cheriguene, S., Dey, N., & Ashour, A. S.: Robust feature selection algorithm based on transductive SVM wrapper and genetic algorithm: application on computer-aided glaucoma classification. IJISTA, 17(3), p.310-346, 2018

[82] Kolar, R., & Jan, J.: Detection of Glaucomatous Eye via Color Fundus Images Using Fractal Dimensions. Radioengineering, 17(3), p.109-114, 2008

[83] Noronha, K. P., Acharya, U. R., Nayak, K. P., Martis, R. J., & Bhandary, S. V.: Automated classification of glaucoma stages using higher order cumulant features. Biomedical Signal Processing and Control, 10, p.174–183, 2014

[84] Mookiah, M. R. K., Acharya, U. R., Lim, C. M., Petznick, A., & Suri, J. S.: Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features. Knowledge-Based Systems, 33, p.73–82, 2012

[85] Dua, S., Acharya, U.R., Chowriappa, P., & Sree, S.V.: Wavelet-based energy features for glaucomatous image classification. IEEE Transactions on Information Technology in Biomedicine, 16(1), p.80–87, 2012

[86] Maheshwari, S., Pachori, R.B., & Acharya, U.R.: Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images. IEEE journal of biomedical and health informatics, 21(3), p.803–813, 2017

[87] Acharya, U.R., Bhat, S., Koh, J.E.W., Bhandary, S.V., & Adeli, H.: A novel algorithm to detect glaucoma risk using texton and local configuration pattern features extracted from fundus images. Computers in Biology and Medicine, 88, p.72-8, 2017

[88] Bock, R., Meier, J., Ny´ul, L. G., Hornegger, J., & Michelson, G.: Glaucoma risk index: automated glaucoma detection from color fundus images. Medical image analysis, 14(3), p.471–481, 2010

[89] Raghavendra, U., Bhandary, S. V., Gudigar, A., & Acharya, U. R.: Novel expert system for glaucoma identification using non-parametric spatial envelope energy spectrum with fundus images. Biocybernetics and Biomedical Engineering, 38(1), p.70-180, 2018

[90] Kumbhare, P., Turkar, M., & Kularkar, R.: Computer Aided Automatic Glaucoma Diagnosis. International Journal of Electrical, Electronics and Data Communication, 2, p.28-32, 2014

[91] Singh, A., Dutta, M. K., ParthaSarathi, M., Uher, V., & Burge, R.: Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. Computer Methods and Programs in Biomedicine, 124, p.108-120, 2016

[92] Maheshwari, S., Pachori, R. B., Kanhangad, V., Bhandary, S. V., & Acharya, U. R.: Iterative variational mode decomposition based automated detection of glaucoma using fundus images. Computers in Biology and Medicine, 88, p.142-149, 2017

[93] Kausu, T. R., Gopi, V. P., Wahid, K. A., Doma, W., & Niwas, S. I.: Combination of clinical and multiresolution features for glaucoma detection and its classification using fundus images. Biocybernetics and Biomedical Engineering, 38(2), p.329–341, 2018

[94] Anushikha, S., Malay, K. D., Partha, S. M., Vaclav, U., & Radim, B.: Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. Computer Methods and Programs in Biomedicine, 124, p.108-120, 2016

[95] Dey, N., Ashour, A. S., & Nguyen, G. N. (2016). Recent Advancement in Multimedia Content using Deep Learning.

[96] Li, Z., Dey, N., Ashour, A. S., Cao, L., Wang, Y., Wang, D., ...& Shi, F.: Convolutional Neural Network Based Clustering and Manifold Learning Method for Diabetic Plantar Pressure Imaging Dataset. Journal of Medical Imaging and Health Informatics, 7(3), p.639-652, 2017

[97] Chen, X., Xu, Y., Wong, D. W. K., Wong, T. Y., & Liu, J.: Glaucoma Detection based on Deep Convolutional Neural Network. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Milan, Italy, 2015

[98] Chai, Y., He, L., Mei, Q., Liu, H., & Xu, L. (2017). Deep Learning Through Two-Branch Convolutional Neuron Network for Glaucoma Diagnosis. In: H. Chen, D. Zeng, E. Karahanna, & I. Bardhan (Eds.) Smart Health. ICSH 2017. Lecture Notes in Computer Science (Vol. 10347, p. 191-201), Springer, Cham.

[99] Zilly, J., Buhmann, J. M., & Mahapatra, D.: Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. Computerized Medical Imaging and Graphics, 55, p.28-41, 2017

[100] Orlando, J. I., Prokofyeva, E., Del Fresno, M., & Blaschko, M. B.: Convolutional neural network transfer for automated glaucoma identification. 12th International Symposium on Medical Information Processing and Analysis, Tandil, Argentina, 2017

[101] Raghavendra, U., Fujita, H., Bhandary, S. V., Gudigar, A., Tan, J. H., & Acharya, U. R.: Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. Information Sciences, 441, p.41-49, 2018

[102] Duerr, B., Haettich, W., Tropf, H., & Winkler, G.: A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals. Pattern Recognition, 12(3), p.189-199, 1980

[103] Chou, Y. Y., & Shapiro, L. G.: A hierarchical multiple classifier learning algorithm. Pattern Analysis and Applications, 6(2), p.150–168, 2003

[104] Azizi, N., Farah, N., & Sellami, M.: Ensemble classifier construction for Arabic handwritten recongnition. 7th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), Tipâza, Algeria, 2011

[105] Brunelli, R., & Falavigna, D.: Person identification using multiple cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(10), p.955-966, 1995

[106] Chibelushi, C. C., Mason, J., & Deravi, F.: Integration of acoustic and visual speech for speaker recognition. Third European Conference on Speech Communication and Technology, Berlin, Germany, 1993

[107] Thakur, N., & Juneja, M.: Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. Biomedical Signal Processing and Control, 42, p.162-189, 2018

[108] Liu, M., Zhang, D., & Chen, S.: Attribute relation learning for zero-shot classification. Neurocomputing, 139, p.34-46, 2014

[109] Zhang, D., & Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage, 59, p.895-907, 2012

[110] Benzebouchi, N. E., Azizi, N., & Bouziane, S. E.: Glaucoma Diagnosis Using Cooperative Convolutional Neural Networks. International Journal of Advances in Electronics and Computer Science, 5(1), p.31-36, 2018

[111] Han, M., & Liu, B.: Ensemble of extreme learning machine for remote sensing image classification. Neurocomputing, 149, p.65–70, 2015

[112] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man and Cybernetics, 9(1), p.62-66, 1979

[113] Sezgin, M., & Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging, 13(1), p.146–1, 2004

[114] Maji, P., & Mullins, R.: On the Reduction of Computational Complexity of Deep Convolutional Neural Networks. Entropy, 20(4), 305, 2018

[115] Wen-Jie, W., & Woo, K. M.: Ultrasound Breast Tumor Image Computer-Aided Diagnosis With Texture and Morphological Features. Academic Radiology, 15(7), p.873–880, 2008

[116] Bob, Z., Xingzheng, W., Fakhri, K., Zhimin, Y., & David, Z.: Computerized facial diagnosis using both color and texture features. Information Sciences, 221, p.49–59, 2013

[117] Chang, J. J., Keun, H. K., Young, J. J., & Jinsung, K.: Improving color and shape repeatability of tongue images for diagnosis by using feedback gridlines. European Journal of Integrative Medicine, 6(3), p.328–336, 2014

[118] Azizi, N., Tlili-Guiassa, Y., & Zemmal, N.: A Computer-Aided Diagnosis System for Breast Cancer Combining Features Complementarily and New Scheme of SVM Classifiers Fusion. International Journal of Multimedia and Ubiquitous Engineering, 8(4), p.45–58, 2013

[119] Azizi, N., Zemmal, N., Sellami, M., & Farah, N.: A new Hybrid Method Combining Genetic Algorithm and Support Vector Machine Classifier: Application to CAD system for mammogram images. International Conference of Multimedia Computing and Systems, p.415-420, 2014

[120] Shradhananda, B., Banshidhar, M., & Ratnakar, D.: Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing, 154, p.1–14, 2015

[121] Loris, N., Sheryl, B., Stefano, G., Emanuele, M., & Tonya, B.: A comparison of methods for extracting information from the co-occurrence matrix for sub cellular classification. Expert Systems with Applications, 40(18), p.7457–7467, 2013

[122] Haralick, R. M.: Statistical and structural approaches to texture. Proceedings of the IEEE, IEEE, 67(5), p.786–804, 1979

[123] Manisha, V., & Balasubramanian, R.: Center symmetric local binary co-occurrence pattern for texture, face and bio-medical image retrieval. Journal of Visual Communication and Image Representation, 32, p.224-236, 2015

[124] Mellisa, P. A., Jeklin, H., & Sakka, N.: Mammograms Classification Using Gray-level Co-occurrenceMatrix and Radial Basis Function Neural Network. Procedia Computer Science, 59, p.83-91, 2015

[125] Luxin, Y., Mingzhi, J., Houzhang, F., Hai, L., & Tianxu, Z.: Atmospheric-Turbulence-Degraded Astronomical Image Restoration by Minimizing Second-Order Central Moment. IEEE Geoscience and Remote Sensing Letters, 9(4), p.672-676, 2012

[126] Angshuman, P., Nilotpal, B., & Ananda, S. C.: Digit Recognition from Pressure Sensor Data using Euler Number and Central Moments. International Conference on Communications, Devices and Intelligent Systems (CODIS), IEEE, Kolkata, India, 2012

[127] Hu, M.: Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, 8(2), p.179-187, 1962

[128] Bradley, A. P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), p.1145–1159, 1997

[129] Nayak, J., Acharya, U. R., Bhat, P. S., Shetty, N., & Lim, T. C.: Automated Diagnosis of Glaucoma Using Digital Fundus Images. Journal of Medical Systems, Springer, 33, p.337-346, 2009

[130] Acharya, U. R., Dua, S., Du, X., Sree, V. S., & Chua, K. C.: Automated diagnosis of glaucoma using texture and higher order spectra features. IEEE Transactions on Information Technology in Biomedicine, 15(3), p.449-455, 2011

[131] Deepti, Y., Partha, M. S., & Malay, K. D.: Classification of Glaucoma Based on Texture Features Using Neural Networks. Seventh International Conference on Contemporary Computing, IEEE, Noida, India, p.109–112, 2014

[132] Kevin, P., Noronhaa, U., Rajendra, A. K., Prabhakar, N., & Roshan, J.: Automated classification of glaucoma stages using higher order cumulant features. Biomedical Signal Processing and Control, 10, p.174–183, 2014

[133] Rajendra, A., Lim, W. J., Kevin, P. N., Lim, C. M., & Sulatha, V. B.: Decision support system for the glaucoma using Gabor transformation. Biomedical Signal Processing and Control, 15, p.18–26, 2015

[134] Nagarajan, R., Balachandran, C., Gunaratnam, D., Klistorner, A., & Graham, S.: Neural network model for early detection of glaucoma using multi-focal visual evoked potential (M-VEP). Investigative Ophthalmology and Visual Science, 43(13), p.U1121-U1121, 2002

[135] Ashish, I., Partha, M. S., & Malay, K. D.: An adaptive threshold based image processing technique for improved glaucoma detection and classification. Computer Methods and Programs in Biomedicine, 122(2), p.229-249, 2015

[136] Simonthomas, S., Thulasi, N., & Asharaf, P.: Automated diagnosis of glaucoma using Haralick texture features. IEEE International Conference on Information Communication and Embedded Systems, Chennai, India, 2014

[137] Issac, A., Partha Sarathi, M., & Dutta, M. K.: An adaptive threshold based image processing technique for improved glaucoma detection and classification. Computer Methods and Programs in Biomedicine, 122(2), p.229–244, 2015

[138] Acharya, U. R., Ng, E. Y. K., Lim, W. J. E., Noronha, K. P., Lim, C. M., Nayak, K. P., & Bhandary, S. V.: Decision support system for the glaucoma using Gabor transformation. Biomedical Signal Processing and Control, 15, p.18-26, 2015

[139] Nnamoko, N., & Korkontzelos, I.: Efficient Treatment of Outliers and Class Imbalance for Diabetes Prediction. Artificial Intelligence in Medicine, 104, 101815, 2020

[140] Li, J., Fong, S., Hu, S., Chu, V. W., Wong, R. K., Mohammed, S., & Dey, N. (2017). Rare Event Prediction Using Similarity Majority Under-Sampling Technique. In: Mohamed A., Berry M., Yap B. (eds) Soft Computing in Data Science. SCDS 2017. Communications in Computer and Information Science, vol 788. Springer, Singapore

[141] Ratadiya, P., & Moorthy, R.: Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification, 2019. ArXiv, abs/1909.04826

[142] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, p.321–357, 2002

[143] Japkowicz, N., & Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), p.429–449, 2002

[144] Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog Artif Intell, 5, p.221–232, 2016

[145] Ashraf, M., Zaman, M., & Ahmed, M.: Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data. Procedia Computer Science, 132, p.1021–1040, 2018

[146] Sarwar, A., Ali, M., Manhas, J. & Sharma, V.: Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int. j. inf. tecnol. 12, p.419–428, 2020

[147] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M.: Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst, 8(1), p.1-14, 2020

[148] Mahabub, A.: A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Appl. Sci., 1(12), 1667, 2019

[149] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, 8, p.6516 - 76531, 2020

[150] Nai-arun, N., & Moungmai, R.: Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science, 69, p.132–142, 2015

[151] Abdillah, A., & Suwarno, S.: Diagnosis of Diabetes using Support Vector Machines with Radial Basis Function Kernels. International Journal Of Technology, 7(5), p.849-858, 2016

[152] Maniruzzaman, M., Kumar, N., Menhazul Abedin, M., Shaykhul Islam, M., Suri, H. S., El-Baz, A. S., & Suri, J. S.: Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine, 152, p.23–34, 2017

[153] Ramani, R., Vimala Devi, K. & Ruba Soundar, K.: MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. Soft Comput, 2020. In press

[154] Tama, B.A., & Rhee, K.:Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artif Intell Rev, 51, p.355–370, 2019

[155] Rahman, M. A., LaPierre, N., & Rangwala, H.: Phenotype Prediction from Metagenomic Data Using Clustering and Assembly with Multiple Instance Learning

(CAMIL). IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(3), p.828-840, 2020

[156] Sankar Ganesh P.V., & Sripriya P. (2020). A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset. In: Smys S., Tavares J., Balas V., Iliyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham

[157] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, p.321–357, 2002

[158] Azizi, N., & Farah, N.: From static to dynamic ensemble of classifiers selection: Application to Arabic handwritten recognition. KES Journal, 16(4), p.279-288, 2012

[159] Azizi, N., Farah, N., Sellami, M., & Ennaji, A. (2010). Using diversity in classifier set selection for Arabic handwritten recognition. In N. El Gayar, J. Kittler, & F. Roli (Eds.), Multiple classifier systems (MCS). Lecture notes in computer science (Vol. 5997, p. 235–244). Berlin, Heidelberg: Springer.

[160] Varma, K.V.S.R.P., Rao, A.A., Sita Maha Lakshmi, T., & Nageswara Rao, P.V.: A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering, 40(5), p.1758–1765, 2014

[161] Bozkurt, M.R , Yurtay, N., Yilmaz, Z., & Sertkaya, C.: Comparison of different methods for determining diabetes. Turk J Elec Eng & Comp Sci, 22(4), p.1044–1055, 2014

[162] Singh N., Singh P. (2020). A Stacked Generalization Approach for Diagnosis and Prediction of Type 2 Diabetes Mellitus. In: Behera H., Nayak J., Naik B., Pelusi D. (eds) Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, vol 990. Springer, Singapore

[163] Choubey, D.K., Paul, S.: GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis. Int J Intell Syst and Appl, 8(1), p.49-59, 2016

[164] Iyer, A., Jeyalatha, S., & Sumbaly, R.: Diagnosis of diabetes using classification mining techniques, 2015. arXiv:1502.03774v1

[165] Ramezani, R., Maadi, M., & Khatami, S. M.: A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal, 57(3), p.1883-1891, 2018

[166] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLOS ONE, 12(7): e0179805, 2017

[167] Chen, P., & Pan, C.: Diabetes classification model based on boosting algorithms. BMC Bioinformatics, 19(1):109, 2018

[168] Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L.: Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset. Fuzzy Information and Engineering, 9(3), p.345–357, 2017

[169] Nai-Arun, N., & Sittidech, P.: Ensemble Learning Model for Diabetes Classification. Advanced Materials Research, vol.931-932, p.1427–1431, 2014

[170] Yilmaz, N., Inan, O., & Uzer, M. S.: A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. J Med Syst, 38:48(5), 2014