Classification du cancer du sein par des approches basées sur les systèmes immunitaires artificiels

Rima Daoudi-Dabladji

NNT : 2016SACLE026

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE D'EVRY VAL D'ESSONNE

ECOLE DOCTORALE N° 580
Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Informatique

Par

Mme Rima Daoudi Ep Dabladji

Classification du cancer du sein par des approches basées sur les Systèmes Immunitaires Artificiels

Thèse présentée et soutenue à Evry, le: 28 septembre 2016

Composition du Jury :

M. William Puech	Professeur	Université Montpellier II	Président du jury
M. Frédéric Bouchara	MCF/HDR	Université de Toulon	Rapporteur
M. Lionel. Fillatre	Professeur	Université de Nice Sophia-Antipolis	Rapporteur
M. Hichem Maaref	Professeur	Université d'Evry Val d'Essonne	Examinateur
M. Khalifa Djemal	MCF/HDR	Université d'Evry Val d'Essonne	Directeur de thèse

Remerciements

Mes premiers remerciements vont tout naturellement à mon directeur de thèse, Mr Khalifa Djemal, Maître de conférences HDR à l'université d'Évry Val d'Essonne. Je désire lui témoigner toute ma reconnaissance pour la confiance qu'il m'a accordée et je souhaite que ce travail soit à la hauteur de ses espérances. Je le remercie également pour son encadrement exemplaire, son accessibilité et sa grande disponibilité.

Je souhaiterai également exprimer ma gratitude à Mr Lionel FILLATRE professeur à l'université de Nice Sophia-Antipolis, et Mr Frédéric BOUCHARA maître de conférences HDR à l'université de Toulon, pour l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de ce modeste travail, dont les remarques et suggestions permettront d'améliorer la qualité de ce manuscrit. Je remercie également Mr Hicham Maaref et Mr Willial Puech respectivement professeurs à l'université d'Évry Val d'Essonne et université de Montpellier II d'avoir accepté d'être membres du jury qui a examiné mon travail.

Je remercie également mes amis et collègues du laboratoire IBISC et de l'UFRST en particulier Khouloud, Pierre Marie, Dalil et Hicham.

Mes vifs et chaleureux remerciements vont aussi à l'endroit de ma famille surtout ma mère qui a consentit tant d'effort matériel et moral pour me soutenir pendant toutes ces longues années d'études. Même si la vie n'a pas toujours été facile dans le foyer familial, tu as su parfois te sacrifier pour nous offrir une vie meilleure. Et grâce à toi, je profite aujourd'hui du meilleur de la vie. J'aimerais pouvoir te rendre tout l'amour et la dévotion que tu nous as offerts, mais une vie entière n'y suffirait pas. J'espère au moins que cette thèse y contribuera en partie.

Et enfin, je réserve un remercient particulier à mon époux Habib, qui a été omniprésent pendant chaque étape de cette thèse. Il m'a soutenu dans les moments les plus durs, je souhaite par cette occasion, lui exprimer tout l'amour du monde.

À ma mère ...

If we knew what it was we were doing, it would not be called research, would it ?
Albert Einstein (1879 - 1955)

Sommaire

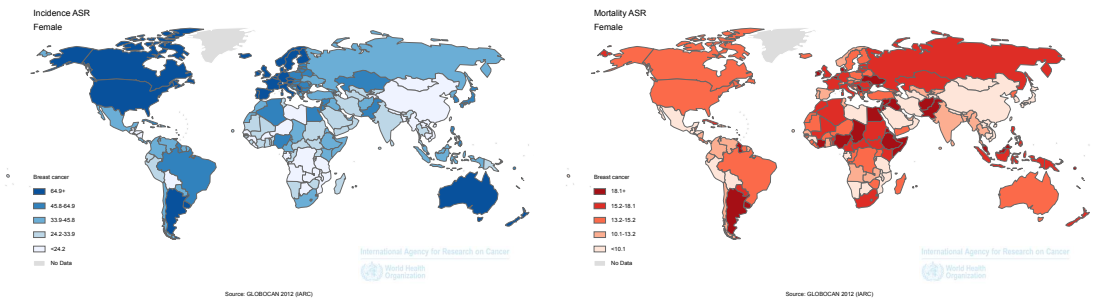
Table des figures

Chapitre 1

INTRODUCTION GENERALE

1.1 Contexte et motivation

Le cancer du sein arrive dans le monde en première position en termes d'incidence et de mortalité parmi les différentes localisations cancéreuses chez les femmes. Selon les chiffres publiés en décembre 2012, par le Centre International de Recherche sur le Cancer (CIRC), qui dépend de l'Organisation Mondiale de la Santé (OMS), l'incidence du cancer du sein a augmenté de plus de 20% entre 2008 et 2012, accompagnée d'une hausse de mortalité de 14%. En effet, 1.7 million de cas sont diagnostiqués chaque année, avec plus de 520 000 décès causés par ce type de cancer [mdlsO13].



(a) (b)

FIGURE 1.1 – Incidence (a) et mortalité (b) du cancer du sein dans le monde selon Globocan [Glo12]

Les symptômes du cancer du sein varient largement, et de nombreux cancers du sein ne présentent aucun symptôme évident, et comme il n'existe pas de moyens de prévention, le type le plus approprié de traitement est le dépistage précoce. En effet, le diagnostic précoce du cancer du sein (maximum 5 ans après la première division cellulaire de la cellule cancéreuse), suivi par un traitement approprié peut réduire le taux de mortalité de 35% environs [Kac12]. Malgré les avancées significatives faites ces dernières décennies en vue d'améliorer la gestion du cancer du sein, des outils de diagnostic plus précis sont encore nécessaires pour aider les oncologues à choisir le traitement nécessaire à des fins de guérison ou de prévention de récidives.

"Il est aujourd'hui urgent, pour mieux lutter contre le cancer, de développer des approches efficaces et abordables pour la détection précoce, le diagnostic et le traitement du cancer du sein chez les femmes vivant dans les pays les moins développés du monde" explique le Dr Christopher Wild, Directeur du CIRC.

Il ne fait aucun doute que l'évaluation et le processus de prise de décision faits par les experts sont des facteurs très importants. Cependant, face au nombre croissant de mammographies effectuées chaque année, grâce aux compagnes de prévention, un temps énorme et une concentration intense sont nécessaires aux radiologues afin de prendre une décision définitive.

Dans ce cadre, des travaux de recherche considérables ont été réalisés dans l'espoir d'apporter de nouvelles perspectives pour l'amélioration du diagnostic du cancer du sein en développant des systèmes de Diagnostic/Décision Assistés par Ordinateur (DAO). Ainsi, l'étude du dépistage du cancer du sein représente un sujet de recherche d'actualité dans l'aide au diagnostic.

Un système automatique de diagnostic assisté par ordinateur (en anglais : (CAD) Computer Aided Diagnosis system) est un domaine en plein croissance de l'analyse d'images médicales qui vise à aider les cliniciens à faire un bon diagnostic. Par exemple, la détection des anomalies, la classification et le diagnostic de ces dernières, et la quantification de la propagation de la maladie [VG10]. De manière générale, les systèmes de DAO servent à donner un second avis au radiologue afin de permettre un diagnostic plus rapide ou plus précis et reproductible. les systèmes de DAO basés sur l'image se composent généralement de quatre modules principaux (figure 1.2) :

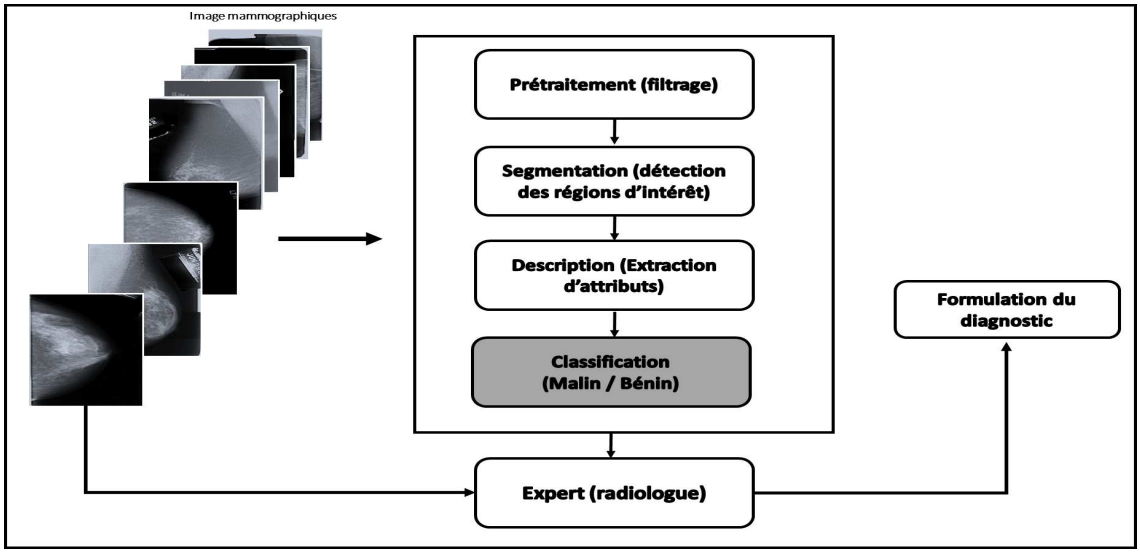


FIGURE 1.2 – Architecture du système de DAO à base d'images mammographiques et de son incorporation dans le diagnostic médical. Adapté par [LLS08]

1. Le prétraitement de l'image vise à améliorer la qualité de l'image en la débruitant et améliorant son contraste. Dans certains cas, le prétraitement peut également gérer la normalisation des données qui permet la comparaison des images obtenues dans des conditions différentes.
2. La segmentation de l'image et la définition des régions d'intérêt (ROI : Regions of interest) cherche à réduire la taille des données analysées en définissant les régions locales où une ou plusieurs lésions peuvent être trouvées. La définition et la segmentation des ROI peuvent être effectuées manuellement, ou d'une manière semi ou entièrement automatique. Dans ce cas, l'analyse ultérieure est réalisée sur les structures segmentées au lieu d'un d'une région définie.
3. La description de l'image porte sur l'extraction des caractéristiques qui décrivent le mieux les images prétraitées à travers des formulations mathématiques. En littérature, la description des images est assurée en utilisant la couleur, la texture et/ou la forme [KDM12].
4. La classification constitue la dernière étape d'un système de diagnostic assisté par ordinateur (DAO). Elle utilise les caractéristiques extraites (résultats de description) pour détecter des anomalies dans les ensembles de données. En règle générale, la classification exige d'abord une étape d'apprentissage où les échantillons d'apprentissage sont fournis de telle sorte qu'une fonction de classification est déduite. En traitement d'images, un échantillon peut représenter un pixel, une zone ou un objet de l'image, ou l'image elle même. La fonction obtenue est ensuite utilisée pour classer les nouveaux échantillons de données. Selon le type de données d'apprentissage, la classification peut être soit supervisée ou non supervisé. L'apprentissage supervisé infère une fonction de classification de données d'apprentissage marqués. Chaque échantillon de données marqué est constitué d'une paire composée d'une entrée (un vecteur de caractéristiques) et une sortie désirée (une étiquette indiquant la normalité ou bien l'anomalie). A l'inverse, l'apprentissage non supervisé ne nécessite pas d'échantillons étiquetés. Ce type d'algorithmes nécessite seulement le vecteur de caractéristiques (attributs) d'entrée pour déduire la fonction de classification.

Le cœur de cette thèse repose sur la dernière étape du système DAO pour le diagnostic du cancer du sein. En effet, une fois que les variables relatives aux masses mammaires extraites, elles doivent être classées comme bénignes/malignes. Le diagnostic du cancer du sein peut donc être défini comme un problème de classification binaire (de deux classes) dans l'apprentissage automatique.

De nos jours une diversité de méthodes sont employées pour aider les médecins dans le diagnostic d'un large éventail de maladies, parmi lesquelles, le cancer du sein. Beaucoup de travaux se sont dirigés vers la détection de la présence de tissus cancéreux dans le sein et la classification de tumeurs. Les approches utilisées proviennent de plusieurs domaines tels que les probabilités et statistiques, les approches connexionnistes ainsi que d'autres outils issus de l'Intelligence Artificielle (IA). En l'occurrence , dans la dernière décennie, un nombre croissant de publications rapportent le progrès des stratégies évolutionnaires

dans la filière du diagnostic et la classification du cancer du sein. Ces algorithmes ont effectivement la capacité de faire évoluer une population de solutions potentielles, telles que les solutions les plus faibles sont éliminées et remplacées par d'autres qui sont meilleurs et incrémentalement plus fortes. En d'autres termes, ces algorithmes suivent le principe de sélection naturelle où chacun à une certaine vraisemblance biologique, et est basé sur l'évolution ou la simulation des systèmes naturels.

Au début du 21ème siècle, une nouvelle technique évolutionnaire a reçu une quantité importante d'intérêt de la part des chercheurs. En effet, l'immunité biologique offre plusieurs caractéristiques attrayantes telles que la mémorisation, l'adaptation à l'environnement, l'apprentissage continu, la prise en compte d'une grande variété de sources d'infections, la sélection automatique, etc. Ces caractéristiques ont encouragé leur adaptation au domaine informatique pour la résolution des problèmes de reconnaissance de formes et d'optimisation. Les chercheurs se sont donc intéressés au système immunitaire biologique en tant que système naturellement doté de mécanismes lui permettant de répondre avec un degré élevé de réactivité et d'adaptabilité aux différentes attaques de pathogènes afin de protéger son hôte.

Le domaine des Systèmes Immunitaires Artificiels (SIA) a vu le jour au milieu des années 80, cependant, c'est seulement dans les années 90 qu'il est devenu un sujet à part entière. Depuis, beaucoup de travaux se sont dirigés vers ce nouveau paradigme de recherche afin de contribuer à la classification des anomalies du sein. On trouve effectivement dans la littérature une variété d'applications qui présentent les SIA comme outil prometteur dans les systèmes DAO. Toutefois, il reste beaucoup d'avantages du Système Immunitaire Naturel (SIN) qui n'ont pas encore été exploités, et qui peuvent améliorer les performances des SIA, tels que la réponse secondaire spécifique, ou la diversité du système qui permet à un antigène d'être reconnu par différentes cellules mémoires.

En tant que tel, cette thèse à pour objectif d'étudier le comportement et les performances des systèmes immunitaires artificiels en termes de précision et de coûts de calcul, sur une tache de classification liée à la détection du cancer du sein. Nous nous focaliserons sur une branche de cette famille d'algorithmes qui est la sélection clonale, et nous mettrons en évidence les qualités et les défauts respectifs des différents outils et méthodes développées et employées.

Comme cela est souvent le cas dans la littérature, la classification des masses mammaires est effectuée en utilisant des bases de données dédiées spécialement à cet effet, où les étapes de prétraitement, description et d'extraction sont effectuées au préalable offrant aux chercheurs des données étiquetées pour le développement des approches de classification. Dans notre travail, les approches proposées sont implémentées et testées sur deux différentes bases de données, très réputées dans la branche de classification du cancer du sein : la Wisconsin Diagnostic Breast Cancer (WDBC) et Digital Database for Screening Mammography (DDSM). Une description plus détaillée de chacune de ces bases de données est donnée dans les annexes A et B respectivement.

1.2 Contribution et organisation du manuscrit

Nous nous sommes intéressés dans cette thèse aux approches de classification des Systèmes Immunitaires Artificiels, et plus précisément aux algorithmes de sélection clonale artificielle, afin d'offrir aux experts un outil performant qui leur servira pour un second avis sur le diagnostic des masses mammaires. Pour mener à bien cette étude, le présent travail est organisé en cinq chapitre dont le premier est l'introduction générale.

Dans le second chapitre on présente une vue générale du Système Immunitaire Naturel, en considérant son anatomie, et le rôle des diverses cellules et organes durant une réponse immunitaire. Les principes de base du système immunitaire ainsi que les mécanismes qui ont inspiré la conception et l'évolution des SIA y sont également abordés. En outre, le chapitre présente aussi un aperçu des trois principales théories immunitaires qui regroupent les différents algorithmes immunitaires artificiels et trace un état de l'art de chacun de ces modèles, à savoir la sélection négative, les réseaux immunitaires et la sélection clonale.

On se focalise dans le chapitre 3 sur le principe de sélection clonale artificielle, qui constitue l'une des caractéristiques les plus importantes de la réponse immunitaire à une stimulation antigénique, afin d'apporter des améliorations à l'algorithme CLONALG. On aborde en premier lieu le principe de ce dernier qui est basé sur un cycle répété de sélection, clonage, mutation et remplacement, en discutant les différentes observations que l'on a pu faire sur son processus d'apprentissage. La suite du chapitre apporte trois différentes contributions qui visent le traitement de ces observations et le renforcement de l'apprentissage de CLONALG.

Dans le chapitre 4, une approche d'optimisation de l'apprentissage des algorithmes de sélection clonale est présentée. Le principe de catégorisation locale des données d'apprentissage est introduit afin de réduire les coûts de calcul de ces algorithmes et améliorer davantage leurs performances. L'approche proposée utilise l'algorithme K-means pour la catégorisation et le réseau de neurones RBF pour l'apprentissage des catégories. Cette procédure permet la réduction du nombre de tests effectués par chaque exemple d'apprentissage afin de sélectionner la cellule à cloner.

Le chapitre 5 est consacré à l'optimisation multimodale par les systèmes immunitaires artificiels. En effet, après avoir étudié la performance des algorithmes de sélection clonale dans le cadre de l'apprentissage automatique, nous nous sommes orientés vers l'exploration de ces derniers, et l'étude de leurs avantages dans le domaine de l'optimisation de fonctions. Le chapitre présente d'abord le principe de l'optimisation et fournie une étude bibliographique sur quelques travaux qui utilisent la sélection clonale artificielle pour traiter ces problèmes. Puis, une approche d'accélération de la convergence du réseau de neurones Perceptron Multi-Couches (PMC) par les SIA est présentée. Le chapitre se termine par une étude comparative avec une optimisation du PMC par un algorithme génétique.

La conclusion de ce travail de recherche, et les directions pour les travaux futurs sont présentés à la fin du manuscrit.

Chapitre 2

IMMUNITÉ BIOLOGIQUE, IMMUNITÉ ARTIFICIELLE ET APPLICATIONS

2.1 Introduction

Le système immunitaire est fortement distribué, très adaptatif, auto-organisé dans la nature, maintient une mémoire des rencontres passées et a la capacité d'apprendre sans cesse de nouvelles rencontres. D'un point de vue informatique, le système immunitaire a beaucoup à offrir à titre d'inspiration pour les chercheurs.

À travers ce chapitre, nous allons d'abord présenter ce qu'est le système immunitaire, son rôle et quelles sont les propriétés fondamentales qui en font une modélisation informatique particulièrement avantageuse.

Nous fournirons par la suite un aperçu des trois principales théories immunitaires qui ont agi en tant que source d'inspiration, à savoir la sélection négative, les réseaux immunitaires et la sélection clonale. On présentera le contexte biologique, l'abstraction artificielle, ainsi qu'une revue bibliographique des différentes applications de la littérature de chacun de ces modèles.

2.2 Brève histoire de l'immunologie

Le terme Immunologie est dérivé du latin Immunis, signifiant déchargé de fardeau où le mot fardeau désigne un impôt ou l'astreinte à une loi [Tab08]. Appliqué à la médecine, il désigne l'état de protection spécifique d'une maladie conféré aux survirants d'une épidémie.

L'immunologie est vieille, au moins d'un demi-milliard d'années pour immunologie adaptative et beaucoup plus longtemps pour l'immunité innée. L'étude de l'immunologie est cependant novice, avec une histoire d'étude scientifique qui remonte à seulement quelques centaines d'années, bien qu'une compréhension empirique ou phénoménologique de cette dernière atteigne probablement loin dans l'antiquité. Aujourd'hui, la science, qui se manifeste dans la discipline de l'immunologie, peut décrire ces phénomènes de façon

plus détaillée. L'immunité est intimement liée à la maladie, elle se réfère à l'état de protection de l'individu vis-à-vis d'agressions étrangères notamment infectieuses. Mais aussi les maladies auto-immunes, et les maladies génétiques multifactorielles, le cancer et les allergies. [Flo08]. Actuellement on préfère une définition plus large, qui considère l'immunologie comme la science de la discrimination du soi (self) et du non-soi (non-self). L'immunité peut donc être définie comme l'ensemble des mécanismes biologiques permettant à un organisme pluricellulaire de maintenir la cohérence de ses cellules et tissus, et d'assurer son intégrité en éliminant ses propres constituants altérés et les substances étrangères auxquelles il est exposé [Pai81].

2.3 Système Immunitaire Naturel

Le système immunitaire Naturel (SIN) est un système de protection complexe mais très adaptable. En cas d'attaque, il défend le corps en détruisant les micro-organismes (les microbes) et les cellules cancéreuses. Composé d'une collection polyvalente et très élaborée de cellules et de molécules qui font équipe pour éliminer un ensemble infini d'envahisseurs. Il s'appuie sur un réseau de communication dynamique très sophistiqué qui transmet les messages entre les différents types de cellules immunitaires.

Le système immunitaire remplit deux fonctions principales : reconnaissance et réponse. Ses cellules reconnaissent tout ce qui ne constitue pas un élément normal du corps. La substance étrangère, aussi appelée antigène, peut être un agent infectieux, un cancer ou un organe transplanté. Une fois l'envahisseur identifié, le corps développe rapidement une contre-attaque pour l'éliminer ou le neutraliser. Le système immunitaire mémorise l'information propre à l'envahisseur, et si celui-ci rapplique, il répondra avec force et rapidité pour l'éliminer.

On appelle réponse immunitaire l'activation des mécanismes du système immunitaire face à une agression de l'organisme. Autrement dit, c'est la production d'anticorps suite à l'introduction d'antigènes dans l'organisme dans le but d'éliminer ou minimiser les dommages qu'ils peuvent induire par la lutte contre les antigènes. Il existe deux grands volets de la réponse immunitaire :

- La réponse non spécifique, qui constitue « l'immunité innée » (nommée ainsi parce qu'elle est présente dès la naissance), agit en ne tenant pas compte de la nature du micro-organisme qu'elle combat ;
- La réponse spécifique, qui confère « l'immunité acquise », passe par la reconnaissance de l'agent à attaquer et la mise en mémoire de cet événement. [DCVZ99]

2.3.1 Immunité innée et immunité spécifique

L'immunité innée (naturelle) est présente à la naissance, elle est appelée ainsi parce que le corps est né avec la capacité de reconnaître certains microbes et immédiatement les détruire. Notre système immunitaire inné, peut détruire de nombreux agents pathogènes dès la première rencontre. L'immunité innée est basée sur un ensemble de récepteurs codés dans les centres germinaux connus sous le nom de récepteurs de reconnaissance des formes

(*P*attern *R*ecognition *R*eceptors : PRR). Ces PRR reconnaissent des motifs moléculaires conservés à la surface de nombreux différents pathogènes : c'est pourquoi on dit que cette immunité est non spécifique.

Bien que non spécifique, la réponse immunitaire innée présente l'avantage d'avoir à sa disposition de façon immédiate un grand nombre de cellules portant les PRR, et donc prêtes à combattre le pathogène, et ce, sans l'avoir rencontré auparavant. Du fait de ces modalités, la réponse innée ne développera pas de mémoire vis à vis du pathogène. De plus, la réponse innée sera comparable lors des différentes rencontres de l'organisme avec le même pathogène. L'immunité naturelle n'est pas toujours suffisante pour éradiquer le pathogène, mais elle est indispensable pour mener à bien une première défense en attendant que l'immunité adaptative ne prenne le relai (4 à 5 jours) [Duf09] [Pen14].L'aspect le plus important de la reconnaissance immunitaire innée réside dans le fait qu'elle induit l'expression de signaux de co-stimulation dans les cellules présentatrices d'antigène, qui mènent à l'activation des lymphocytes T, favorisant le début de la réponse immunitaire adaptative.

L'immunité acquise (adaptative ou spécifique), n'est pas présente à la naissance, elle est apprise. Pendant que le système immunitaire d'une personne rencontre des corps étrangers (antigènes), les composants de l'immunité acquise apprennent la meilleure façon d'attaquer chaque antigène, et commencent à développer une mémoire pour cet antigène.

L'immunité acquise est aussi appelée immunité spécifique, car elle adapte son attaque à un antigène spécifique rencontré précédemment. Ses caractéristiques sont sa capacité à apprendre, s'adapter, et se souvenir. L'immunité acquise prend du temps à se développer après la première exposition à un nouvel antigène. Cependant, après la mémorisation de la forme l'antigène, les réponses suivantes à ce denier sont plus rapides et plus efficaces que celles qui ont eu lieu avant la première exposition. Les acteurs principaux de l'immunité acquise sont les lymphocytes (cellules T et B), et les immunoglobulines. Le rôle de l'immunité adaptative est de détruire les agents pathogènes envahisseurs, et toutes les molécules toxiques qu'ils produisent. Parce que les réponses adaptatives sont destructrices, il est essentiel qu'elles soient faites uniquement en réponse à des molécules étrangères au corps humain. La capacité de distinguer ce qui est étranger de ce qui est auto est une caractéristique fondamentale du système immunitaire adaptatif. De temps en temps, le système ne parvient pas à faire cette distinction et réagit contre ses propres molécules. Ces maladies auto-immunes peuvent être fatales [AB12].

2.3.2 Concepts immunologiques

Le système immunitaire étant très complexe, doit être vu sous différents angles. Il peut être considéré en tant qu'un ensemble d'organes, cellules, et molécules distribués dans tout le corps.

2.3.2.1 Organes immunitaires

Les organes immunitaires, et leurs principales fonctions, comprennent (voir la figure 2.1) :

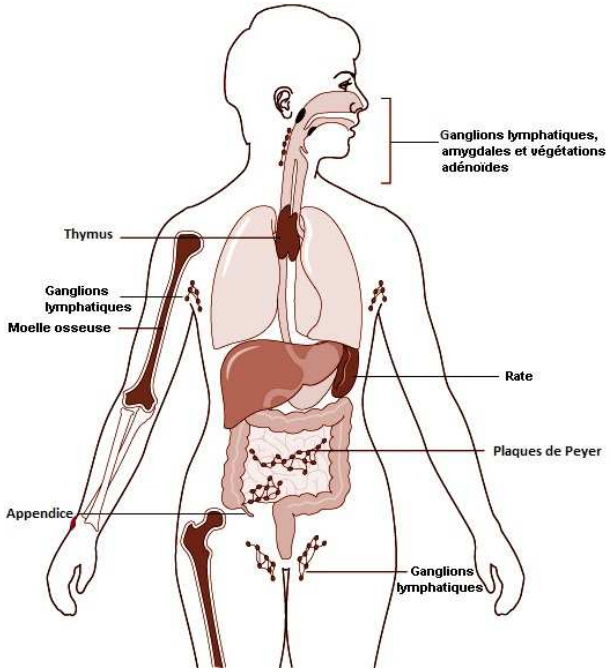


FIGURE 2.1 – Anatomie du système immunitaire naturel, les organes immunitaires [Ste98].

— Amygdales et adénoïdes : ganglions lymphatiques spécialisés contenant les cellules immunitaires qui protègent le corps contre des envahisseurs de l'appareil respiratoire ;

— Thymus : quelques cellules migrent dans le thymus, de la moelle osseuse, où elles se multiplient et mûrissent, se transformant en cellules T, capables de produire une réponse immunitaire ;

— Ganglions lymphatiques : servent de sites de convergence des vaisseaux lymphatiques, où chaque nœud stocke les cellules immunitaires, y compris les cellules B et T. Site où la réponse immunitaire adaptative a lieu ;

— Rate : site où les leucocytes détruisent les organismes qui envahissent la circulation sanguine ;

— Appendice et les plaques de Peyer : ganglions lymphatiques spécialisés contenant les cellules immunitaires destinées à protéger le système digestif.

— Moelle osseuse : les tissus mous contenus dans la partie intérieure des plus longs os, responsables de la génération des cellules immunitaires. C'est le lieu de maturation des lymphocytes B.

2.3.2.2 Cellules immunitaires

Les cellules impliquées dans la défense de l'organisme sont les globules blancs, aussi appelées leucocytes. Elles sont produites dans la moelle osseuse, où beaucoup d'entre elles

mûrissent et migrent vers les tissus en circulant dans les vaisseaux sanguins et lymphatiques. Certaines d'entre elles sont responsables de la défense générale, tandis que d'autres sont formées pour lutter contre les agents pathogènes hautement spécifiques. Pour un fonctionnement efficace, une coopération continue entre ces cellules est nécessaire. Il existe différents types de globules blancs, les plus important étant les lymphocytes.

Les lymphocytes sont des petits leucocytes (globules blancs) qui possèdent une grande responsabilité dans le système immunitaire. Il existe deux principaux types de lymphocytes : lymphocytes B (ou cellules B), qui lors de l'activation, se différencient en plasmocytes (ou des cellules plasmatiques) capables de sécréter des anticorps ; et les lymphocytes T (ou cellules T).

a) **Lymphocytes B :** les cellules B sont des cellules participantes à la réponse immunitaire spécifique, leurs principales fonctions comprennent la production et la sécrétion d'anticorps (Ac) en réponse à des protéines exogènes tels que les bactéries, les virus et les cellules tumorales. Chaque cellule est programmée pour produire un anticorps spécifique. Les anticorps sont des protéines spécifiques qui reconnaissent et se lient à une autre protéine particulière.

   Il existe plusieurs types des cellules B : les cellules B *immatures ou naïves*, qui n'ont encore jamais rencontré leur antigène de prédilection. Après rencontre de l'antigène et activation, les lymphocytes B se multiplient en plusieurs clones, une partie se différencie en *plasmocytes* qui secrètent des immunoglobulines (anticorps), en vue de la destruction des antigènes. Le reste donne des *lymphocytes B mémoires*, qui vivent plus longtemps que les plasmocytes. Elles ont pour rôle de mémoriser les propriétés de l'antigène les ayant activées, afin de créer une réponse immunitaire plus rapide, plus intense et plus spécifique, dans le cas d'une seconde représentation du même antigène (réponse immunitaire secondaire). [SA13] [Let07] .

b) **Lymphocytes T :** Les cellules T sont appelées ainsi parce qu'elles arrivent à maturité dans le thymus [Dre95]. Elles ont pour fonction la réglementation des actions des autres cellules, et attaquer directement les cellules hôtes infectées. Les lymphocytes T peuvent être subdivisés en trois sous-classes principales : les cellules T aideuses, (T helper), les cytotoxiques ou cellules T tueuses (T killer) et les lymphocites T suppresseurs (suppressor T cells).

   Les cellules T *aideuses* sont essentiellement chargées de l'activation des cellules B. Les cellules T *tueuses*, sont capables d'éliminer les envahisseurs microbiens, les virus ou les cellules cancéreuses. Une fois activées et liées à leurs ligands, elles injectent des produits toxiques dans les autres cellules, perforant leur membrane de surface, et provoquant leur destruction. Les lymphocytes T *suppresseurs* quant à elles, sont essentielles au maintien de la la réponse immunitaire. Elles servent à éviter les réactions immunitaires non appropriées (maladies auto-immunes) [JTWC05].

   Les lymphocytes B et T expriment sur leur surfaces, des récepteurs hautement spécifiques pour un déterminant antigénique donné. Le récepteur des cellules B est une forme de la molécule d'anticorps lié à la membrane, et qui sera sécrété après que la cellule soit convenablement activée.

c) **Les cellules tueuses naturelles :** (natural killer cells) constituent un autre type de lymphocytes mortels. Comme les lymphocytes T tueuses, elles contiennent des granules

remplis de produits chimiques puissants. Elles sont désignées tueuses naturelles parce qu'à la différence des cellules T tueuses, elles ne nécessitent pas de reconnaître un antigène spécifique avant qu'elles ne commencent à agir. Elles attaquent principalement les tumeurs et protègent contre une grande variété de microbes infectieux. Ces cellules contribuent également à la régulation du système immunitaire, sécrétant de grandes quantités de lymphokines.

2.3.2.3 Les Antigènes

Un antigène est une molécule ou un micro-organisme d'origine biologique ou synthétique , classiquement réputé étranger à l'organisme, et capable d'engendrer une réponse immunitaire. Une fois pénétré dans l'organisme, l'antigène n'est pas détecté dans sa totalité par le système immunitaire, ce dernier ne détecte que des déterminants antigéniques, appelés épitopes, sur l'antigène. Ces épitopes se lient de manière complémentaire avec les paratopes des anticorps. Un même antigène peut comporter plusieurs épitopes (identiques ou différents), ainsi un antigène induit la synthèse de plusieurs différents anticorps (un pour chaque épitope différent). Si ses épitopes sont reconnus comme appartenant au non-soi, alors il est lui-même immédiatement reconnu comme appartenant au non-soi. La reconnaissance épitope/paratope constitue la base de la réponse immunitaire spécifique permettant la sélection clonale, i.e, la sélection des cellules capables de s'attaquer spécifiquement à l'antigène correspondant à un épitope particulier.

2.3.2.4 Les Anticorps

Les anticorps sont des molécules complexes appartenant à la famille des immunoglobulines (ce qui explique que l'abréviation courante d'anticorps soit Ig). Fabriqués par les plasmocytes des lymphocytes B activés, la fonction unique des anticorps est de reconnaître et se fixer de façon spécifique aux antigènes.

À travers la reconnaissance et la distinction de motifs moléculaires spécifiques, les anticorps jouent un rôle central dans le système immunitaire. Les antigènes sont diversifiés dans leur structure, forçant le répertoire d'anticorps d'être de grande taille [Ton83].

• Structure d'un anticorps :

L'unité de base d'un anticorps (ou immunoglobulines), est composée de quatre chaînes protéiques identiques deux à deux : deux chaînes légères (L pour light), et deux chaînes lourdes (H pour heavy), en fonction de leurs poids moléculaires respectifs. Les chaînes sont disposées en une forme caractéristique de Y, et maintenues entre elles par des ponts disulfures. Chacune des chaînes légères est constituée d'une région variable "V" (car elle est différente pour chaque anticorps) principalement responsable de la reconnaissance de l'antigène, et d'une région constante (C) responsable de diverses fonctions effectrices, telles que l'activation du système immunitaire ou la fixation de l'anticorps sur les surfaces (voir la figure 2.2).

La partie de l'anticorps qui se combine à l'épitope de l'antigène est appelée paratope. Le paratope est formé par une combinaison entre un domaine variable porté par une chaîne lourde, et t le domaine variable adjacent porté par une chaîne légère de l'anticorps.

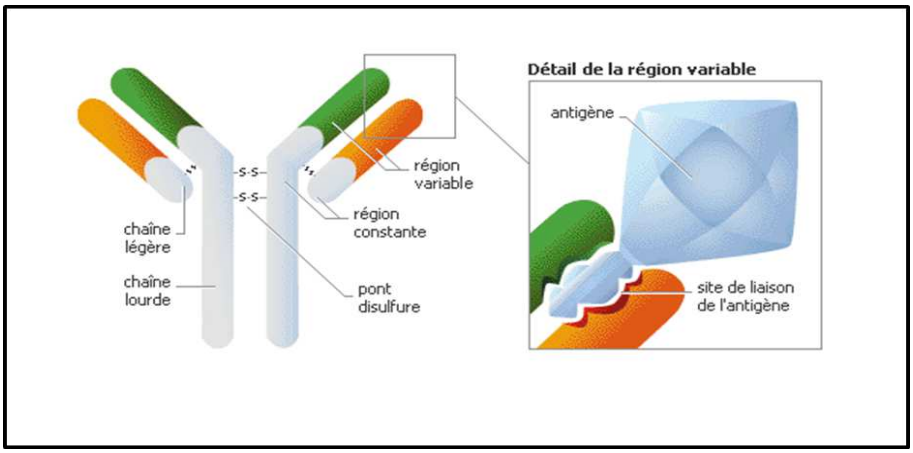


FIGURE 2.2 – Structure d'un anticorps [Enc].

Un anticorps possède ainsi deux paratopes identiques, lui permettant de se lier à deux molécules d'antigène.

Le complexe épitope/paratope, est au final un peu comme un puzzle (figure 2.3), le but des 2 pièces étant de s'emboîter l'une dans l'autre lorsqu'elles reconnaissent des séquences polypeptidiques spécifiques. Ainsi, dés qu'un épitope est détecté, le paratope peut agir et déclencher les signaux appropriés pour activer la réponse immunitaire. Le paramètre désignant la force de la liaison entre le paratope et l'épitope est appelé affinité.

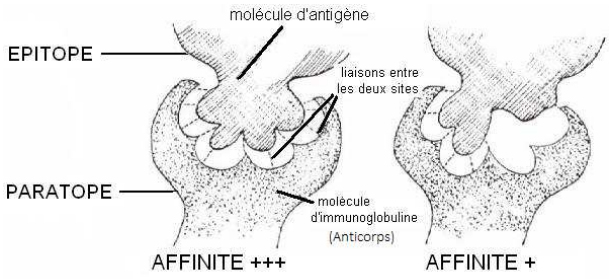


FIGURE 2.3 – Liaison épitope / paratope [Dec04].

2.3.2.5 Architecture du système immunitaire

Le système immunitaire a une architecture multi-couches, avec des défenses sur plusieurs niveaux (voir figure 2.4).

• Barrières physiques : notre peau fonctionne comme un bouclier pour la protection du corps contre les envahisseurs. Le système respiratoire contribue également à maintenir loin les antigènes. La peau et les muqueuses qui tapissent les voies respiratoires et digestives contiennent également des anticorps.

• Barrières physiologiques : les fluides tels que la salive, la sueur et les larmes contiennent des enzymes destructrices. les acides de l'estomac tuent la plupart des microorganismes

ingérés dans les aliments et l'eau. Le pH et la température du corps présentent aussi des conditions de vie défavorables pour certains envahisseurs.

• L'immunité innée et l'immunité adaptative constituent aussi des couches du système immunitaire (voir section 2.3.1).

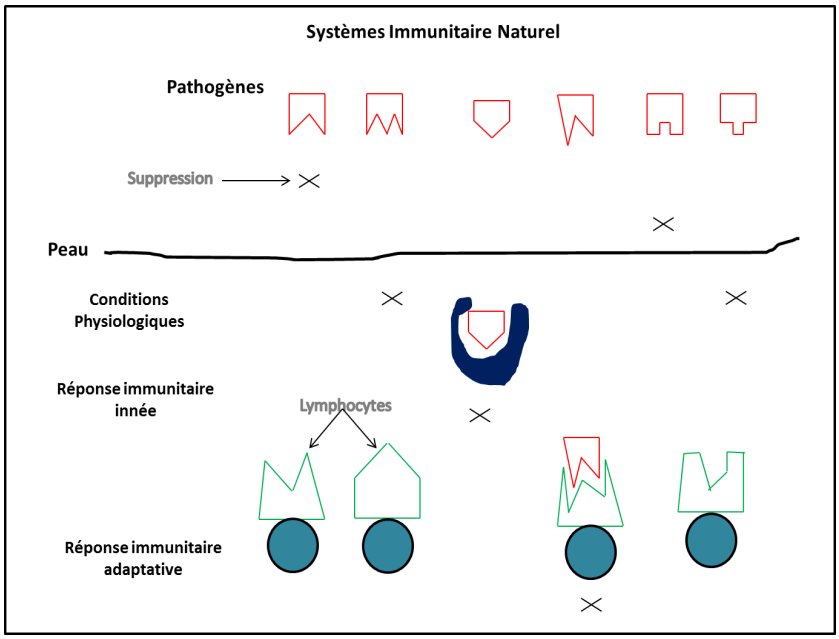


FIGURE 2.4 – Structure multi-couche du système immunitaire.

2.3.3 Déroulement de la réponse immunitaire

Quand un antigène pénètre dans le corps, une réponse immunitaire innée est lancée en première ligne de défense. Les phagocytes (cellules du SI inné) détruisent partiellement l'antigène, puis le présentent comme des fragments de peptides antigéniques liés à leurs surface, afin qu'il soit reconnu par les lymphocytes. Une fois que les cellules T *reconnaissent* l'antigène, elles commencent à produire et à secréter les cytokines (signaux chimiques), agissant sur les lymphocytes B en vue de les mobiliser. Dès leur *activation*, les cellules B qui identifient l'antigène avec un certain degré d'affinité, se multiplient grâce au *clonage*. Certaines se différencient en cellules mémoires *(mémorisation)* sensibilisées à cet antigène, d'autres se transforment en cellules sécrétrices d'anticorps. Un grand volume d'anticorps est produit et libéré de la surface de ces cellules pour faire face à l'antigène et le *neutraliser*. L'antigène recouvert d'anticorps est ensuite reconnu par les phagocytes, et éliminé.

Lors du clonage, les cellules subissent une *maturation de l'affinité*, les récepteurs antigéniques sont mutés, et les fils qui se lient à l'antigène avec le meilleur degré d'affinité, réussissent à survivre. La maturation de l'affinité (clonage + mutation) est un mécanisme important de la réponse immunitaire qui vise à améliorer la réponse aux antigènes, et la création de cellules mémoires ayant une longue durée de vie, qui serviront à répondre plus

rapidement aux futurs attaques.

Lors du premier contact avec un antigène, la réponse immunitaire est dite **primaire**, elle est caractérisée par la production d'anticorps et de cellules mémoires , et se développe en plusieurs jours. Au deuxième contact avec le même antigène, les cellules mémoires produites sont responsables de la réponse immunitaire **secondaire**, plus rapide et plus intense (figure 2.5).

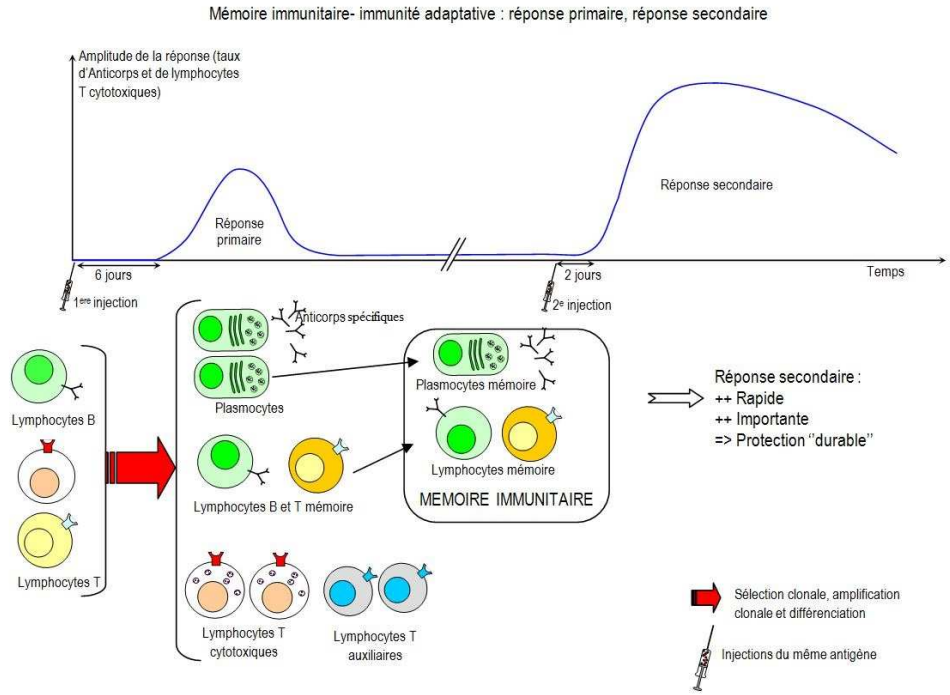


FIGURE 2.5 – Réponse immunitaire primaire VS réponse immunitaire secondaire [MV].

2.4 Systèmes Immunitaires Artificiels

Comme nous l'avons vu dans les sections précédentes, le système immunitaire est un système très complexe qui engage une multitude de tâches. Les capacités du système immunitaire naturel ont contribué à inspirer les chercheurs en informatique à établir des systèmes immunitaires artificiels. d'une certaine façon, ces systèmes imitent certaines propriétés informatiquement attrayantes du système immunitaire naturel, dans le but de construire des solutions plus robustes et adaptables.

Le domaine des Systèmes Immunitaires Artificiels (SIA) porte d'une part sur l'abstraction de la structure et des fonctions du système immunitaire biologique à des systèmes informatiques. d'autre part, il permet et d'enquêter sur l'application de ces systèmes en vue de résoudre des problèmes du monde réel et d'ingénierie. Les SIA sont un sous-domaine de l'informatique bio-inspirée, et du calcul naturel, avec des intérêts pour l'apprentissage

automatique. C'est une branche de l'intelligence artificielle qui constitue une famille d'algorithmes intelligents qui ont été définis par de Castro & Timmis [DCT02c] comme :

« des systèmes adaptatifs, inspirés par l'immunologie théorique ainsi que par les fonctions, principes et modèles immunitaires observés, qui sont appliqués à la résolution des problèmes ».

Dans cette partie du chapitre, nous explorerons les fondements des SIA, traçant leur bref historique, et décrivant quel type d'immunologie a servi d'inspiration. Nous verrons que différents processus et idées immunitaires ont été capturés et utilisés avec succès dans la reconnaissance de formes, la détection de défauts, le diagnostic, la sécurité informatique, et une variété d'autres applications.

2.4.1 Historique

Les systèmes immunitaires artificiels ont vu le jour par le premier travail d'immunologie théorique de Farmer, Perelson et Varela [FPP86] [Per89] [VCDV88]. Ces travaux ont examiné un certain nombre de modèles de réseaux immunitaires théoriques proposés pour décrire la maintenance de la mémoire immunitaire. Cependant c'est seulement dans le milieu des années 1990 que les SIA devinrent un sujet à part entière. En 1999 Dasgupta a édité le premier livre sur les Systèmes Immunitaires Artificiels [Das99], suivi par les travaux de De Castro & Von Zuben et Nicosia & Cutello sur la sélection clonale (CLONALG) en 2002 [DCVZ02] [CN02]. La branche des systèmes immunitaires artificiels a été largement étudiée au cours des dernières années, avec plusieurs approches et applications dans la littérature. En raison d'un nombre croissant de travaux menés sur les SIA, la conférence internationale sur les systèmes immunitaires artificiels (ICARIS) a été lancée en 2002 et a opéré dans les années suivantes.

2.4.2 Aspects computationnels du système immunitaire

Le système immunitaire naturel est un sujet de grand intérêt pour la recherche, en raison de ses puissantes capacités de traitement de l'information. Notamment, les nombreux calculs complexes qu'il effectue d'une manière hautement parallèle et distribuée. Le système immunitaire fonctionne comme une sorte de «second cerveau», car il peut stocker des souvenirs des expériences antérieures dans les forces des interactions de ses cellules constitutives. Il est également capable de générer des réponses à des modèles nouveaux et novateurs (antigènes). En outre, la réponse immunitaire se développe dans le temps, et la description de son temps d'évolution est un problème intéressant dans les systèmes dynamiques [DF96]. Les principales caractéristiques du système immunitaire qui fournissent plusieurs aspects importants au champ de traitement de l'information peuvent être récapitulées en vertu des termes informatiques suivants :

• **Reconnaissance :** Le système immunitaire peut reconnaître et classer les différents modèles et générer des réponses sélectives. La reconnaissance est obtenue par la liaison intercellulaire. La discrimination entre le soi et le non-soi est l'une des principales tâches que le système immunitaire résout au cours du processus de reconnaissance.

• **Extraction de caractéristiques :** les cellules présentatrices d'antigènes (CPA) interprètent le contexte antigénique, et extraient les caractéristiques par le traitement et la présentation des peptides antigéniques sur leurs surfaces. Chaque CPA détruit le bruit moléculaire, afin de concentrer l'attention des récepteurs des lymphocytes [CCS96].

• **Diversité :** le système immunitaire utilise l'analyse combinatoire [Sak84] (en partie par un procédé génétique) en vue de générer un ensemble varié de récepteurs lymphocytaires, pour faire en sorte qu'au moins certains lymphocytes peuvent se lier à un quelconque antigène donné (connu ou inconnu).

• **Apprentissage :** Il apprend, par expérience, la structure d'un antigène spécifique. Apporter des changements dans la concentration des lymphocytes est le mécanisme d'apprentissage, et il se déroule au cours de la réponse primaire (première rencontre avec l'antigène). Donc, la capacité d'apprentissage du système immunitaire réside principalement dans le mécanisme de recrutement, qui génère de nouvelles cellules du système immunitaire sur la base de l'état actuel du système (aussi appelé expansion clonale).

• **Mémorisation :** Lorsque les lymphocytes sont activés, quelques uns de chaque type deviennent des cellules mémoires spéciales. La longévité inhérente des cellules mémoires immunitaires est dynamique, et nécessite une stimulation continue par les antigènes. Le système immunitaire maintient un équilibre idéal entre l'économie et la performance dans la conservation d'une mémoire minimale mais suffisante du passé.

• **Distribution :** le système immunitaire est naturellement distribué. Les cellules immunitaires, en particulier les lymphocytes, subissent des niveaux constants de recirculation à travers le sang, la lymphe, les organes lymphoïdes et les tissus. Si elles rencontrent des défis antigéniques, elles stimulent les réponses immunitaires spécifiques.

• **Mécanisme de seuil :** la réponse immunitaire et la prolifération (clonage) des cellules immunitaires prend place au-dessus d'un certain seuil d'adaptation (affinité épitope/paratope).

• **Protection dynamique :** l'expansion clonale et l'hypermutation (clonage et mutation) permettent la génération de cellules immunitaires de hautes affinités. Ce processus (également appelé maturation d'affinité), équilibre dynamiquement l'exploration par rapport à l'exploitation dans l'immunité adaptative. La protection dynamique augmente la couverture assurée par le système immunitaire au fil du temps.

• **Parallélisme :** Le système immunitaire est capable de produire plusieurs réponses immunitaires en même temps à des endroits dispersés.

D'autres caractéristiques connexes comme l'adaptabilité, la spécificité, l'auto-tolérance, la différenciation etc. remplissent également des fonctions importantes dans la réponse immunitaire. Toutes ces propriétés remarquables de traitement d'informations du système immunitaire fournissent plusieurs aspects importants dans le domaine computationnel.

2.4.3 Structure de conception d'un SIA

Dans une tentative de création d'un modèle commun pour les SIA, les auteurs de [dCT02a] ont adopté un schema de structure pour la conception d'un algorithme de SIA, nécessitant au moins les éléments de base suivants :

• Une représentation des composants du système (modèles abstraits des cellules immunitaires) ;

• Un ensemble de fonctions pour évaluer l'affinité (la similarité) entre les composants du systèmes ;

• Un ensemble d'algorithmes pour contrôler l'évolution et la dynamique du système immunitaire artificiel.

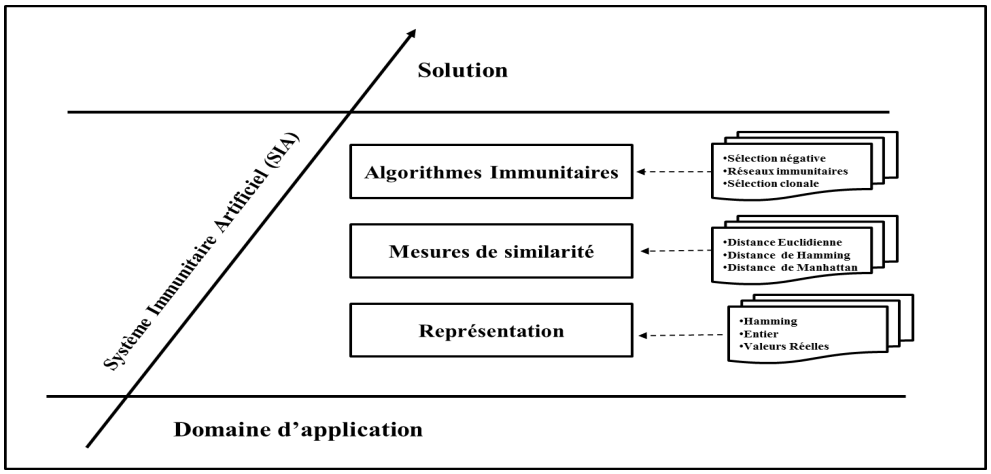


FIGURE 2.6 – Structure de conception d'un système immunitaire artificiel adaptée par [dCT02a].

Cette structure peut être considéré comme une approche multicouches suivant les trois éléments de base pour la conception d'un algorithme d'inspiration biologique (figure 2.6).

— **Représentation :** Afin de construire un système tel qu'un SIA, un domaine d'application ou une fonction cible sont généralement exigés. A partir de cette base, la façon dont les éléments du système (cellules) seront représentés est considérée. Cette façon est appelée espace de forme (shape space) [Per89]. Il existe plusieurs types d'espaces de forme, tels que Hamming, les valeurs réelles, etc. chacun porte son propre biais et doit être choisi avec précaution [FT03].

— **Mesures de similarité :** Une fois la représentation choisie, une ou plusieurs mesures d'affinité sont utilisées pour quantifier les interactions entre les éléments du système. Il y a beaucoup de mesures de similarité possibles (qui sont partiellement dépendantes de la représentation adoptée), le plus souvent on recourt à l'utilisation des métriques de distances comme la distance *Euclidienne*, la distance de *Manhattan* ou la distance de *Hamming* [dCT03]. En intelligence artificielle, et en particulier en classification, l'affinité est appelé similarité.

— **Les algorithmes immunitaires :** La couche finale implique l'utilisation d'algorithmes qui régissent le comportement (dynamique) du système. Ces algorithmes incluent ceux basés sur les processus immunitaires suivants : sélection négative, les réseaux immunitaires, et la sélection clonale. Le principe de chacun de ces algo-

rithmes sera détaillé dans les sections qui suivent.

2.5 Modèles immunitaires de base et principes biologiques

Les principaux développements au sein des systèmes immunitaires artificiels, ont mis l'accent sur trois principales théories immunologiques : la sélection clonale, les réseaux immunitaires et la sélection négative. Les chercheurs dans le domaine des SIA se sont concentrés en majeure partie, sur les mécanismes d'apprentissage et de mémorisation du système immunitaire. Ces mécanismes sont inhérents à la sélection négative et aux réseaux immunitaires, et au principe de sélection clonale. Dans cette section, nous décrivons les trois principales théories immunitaires mentionnées ci-dessus qui ont agi en tant que source d'inspiration, en présentant le contexte biologique, l'abstraction artificielle, ainsi qu'une revue bibliographique des travaux de base de chaque modèle .

2.5.1 Sélection Négative

La sélection négative est un processus de sélection qui a lieu dans le thymus. Elle se produit lorsque les cellules T qui attaquent «soi» sont éliminées, et seulement celles (cellules T) qui répondent faiblement au soi, le cas échéant, sont libérées dans le système lymphatique. Ces cellules répondent alors à des degrés divers, aux différents types d'envahisseurs du «non-soi» qui doivent être retirés de la circulation sanguine, la lymphe, les tissus, etc. Dans la littérature des SIA, ce processus est souvent appelé détection de virus, mais dans l'essentiel, le processus de sélection négative devrait fournir un ensemble de détecteurs (cellules T) permettant de détecter tout changement dans le «soi», ou toute forme de «non-soi» (virus ou autres). C'est ce processus qui a été soustrait pour former un algorithme de détection de changement d'un ensemble prédéfini d'objets.

2.5.1.1 Développement de la Sélection Négative Artificielle

La Sélection Négative Artificielle (SNA) a été introduite en 1994 dans [FPAC94], récemment, elle est souvent appelée détection négative. Le principe de la sélection négative biologique a inspiré les auteurs à proposer un algorithme pour détecter la manipulation de données causée par des virus informatiques. L'idée de base est de générer un certain nombre de détecteurs, puis d'appliquer ces détecteurs pour classer de nouvelles données (invisibles) comme des données du soi ou du non-soi. Ils ont considéré l'algorithme de la sélection négative comme un processus de détection d'anomalies composé de trois phases principales :

— La définition du soi.
— La génération des détecteurs.
— Le contrôle d'occurrence des anomalies.

L'algorithme général de sélection négative, introduit par[FPAC94] est résumé dans les étapes suivantes :

1. Définir un ensemble donné de modèles du soi S,
2. Générer aléatoirement un ensemble de détecteurs R_0,
3. Pour chaque $r0 \in R_0$, créer un ensemble R de detecteurs, qui n'identifient aucun élément appartenant a l'ensemble S de la manière suivante :
   (a) Calculer la similarité entre chaque cellule r_0 et toutes les cellules du soi $s \in S$
   (b) Si la similarité entre un élément r_0 et au moins un élément s est supérieur ou égal a un seuil de similarité prédéfini : $r0$ est considéré comme un élément de soi et supprimé .
   (c) Sinon il sera considéré comme un détecteur de non soi et sera ajouté à l'ensemble de détecteur R.

L'algorithme de la sélection positive [CS92] est une alternative de l'algorithme de la sélection négative. La principale différence est la génération des détecteurs qui permettent la détection des éléments du soi. Selon cet algorithme, un élément de non-soi suspect est comparé avec l'ensemble des détecteurs du soi ; s'il n'est pas détecté alors il est considéré comme un élément de non soi.

2.5.1.2 Étude bibliographique

Les algorithmes de sélection négative ont été utilisés dans différents domaines d'applications, notamment la détection d'anomalies. Dans [ATdL$^+$02] l'algorithme de sélection négative avec mutation (NS Mutation algorithm) a été présenté. En plus d'utiliser la mutation des détecteurs, ce travail permet d'éliminer la redondance et dispose de paramètres ajustables. Il se compose de trois phases : la définition du soi, la génération des détecteurs, et la comparaison des détecteurs générés avec les éléments du soi en se basant sur un seuil de similarité. Les auteurs de [GC04] ont présenté une approche de sélection négative auto-adaptative pour la détection d'anomalies. L'algorithme emploie des techniques auto-adaptatives pour le réglage de paramètres. Les deux principales phases de l'algorithme comprennent la génération de la population initiale de détecteur, et l'évolution de cette dernière. Un algorithme de sélection négative à valeurs réelles avec des détecteurs de taille variable (Real-Valued Negative Selection Algorithm with Variable-Sized Detectors) nommé V-Detector a été développé dans [JD04]. Le travail décrit une approche avec beaucoup de caractéristiques remarquables. Comme une stratégie de génération et un schéma de détection simples. Avec une taille variable des détecteurs,une estimation de la couverture et la technique de connaissance de frontières, pour interpréter l'ensemble des données d'apprentissage dans son ensemble, et non pas en tant que points indépendants. Cet algorithme a été exposé à des tests de sensibilité dans[JD09], afin d'exposer ses avantages et ses inconvénients. Dans [IO09], les auteurs ont proposé un nouvel algorithme de sélection négative nommé Artificial Negative Selection Classifier (ANSC) pour la classification multi-classes. Ce travail introduit une méthode de découpage pour réduire l'effet du bruit. Il combine l'algorithme de sélection négative avec le mécanisme de sélection clonale, pour résoudre les problèmes qui empêchent les algorithmes de sélection négative d'être appliqués à des problèmes de classification. Ces problèmes incluent la recherche aléatoire, le surapprentissage et les données incomplètes. Un autre algorithme de sélection négative basé chaos (Chaotic-based hybrid negative selection algorithm) a

été proposé dans[AKA10]. L'algorithme utilise des cartes chaotiques pour la sélection de paramètres, de sorte que la zone de couverture de l'algorithme de sélection négative doit être augmentée, et la sélection clonale pour obtenir des détecteurs optimaux et non-chevauchants. Ce travail a montré un meilleur résultat pour le problème de la détection d'anomalies. Les auteurs dans[GZMJ12] ont proposé une approche améliorée de sélection négative pour la détection d'anomalies en intégrant une stratégie supplémentaire à l'étape d'apprentissage (extra-training), afin de générer des détecteurs qui couvrent la région du soi. L'objectif principal de cette étape supplémentaire est d'éliminer les exemples du soi, et réduire les taux de calcul dans l'étape d'apprentissage. La couverture de soi a aussi été améliorée par cette dernière. D'autres améliorations de la SNA ont été publiées ([Che13], [ZQT13], [MKS+14]).

2.5.2 Réseaux Immunitaires

Jusqu'à présent, nous avons vu qu'une partie d'un anticorps (connu sous le nom de paratope) va se lier à une partie d'un antigène (connu sous le nom d'épitope). Dans un document historique marquant [Jer74], Jerne va plus loin en proposant le système immunitaire comme un système capable de réaliser une mémoire immunitaire par l'existence d'un réseau de renforcement mutuel des cellules B (anticorps). Le principe théorique est que les anticorps ont aussi des épitopes (appelés idiotopes), qui peuvent être liés par les paratopes d'autres anticorps. La liaison entre idiotopes et paratopes a pour effet de stimuler les anticorps. En effet, les paratopes sur anticorps réagissent aux idiotopes sur des cellules similaires, comme qu'ils le feraient contre un antigène. Cependant, pour contrer cette réaction, il existe une certaine quantité de suppression entre les anticorps qui agit en tant que mécanisme de régulation. Cette interaction entre les anticorps, contribue à une structure de mémoire stable, et compte pour le maintien des cellules de mémoire, même en l'absence d'antigène. Cette théorie a été raffiné et a été officialisée dans des œuvres successives par [FPP86] et [Per89]. En plus, ils ont fait une hypothèse simplificatrice d'une unique région de liaison à l'épitope sur chaque anticorps et chaque antigène, et une région de liaison au paratope unique sur chaque anticorps. Après avoir présenté leur modèle, ils ont également discuté les possibilités d'utilisation de ces descriptions immunologiques abstraites comme des outils informatiques pour l'apprentissage [HR02].

2.5.2.1 Développement des Réseaux Immunitaires Artificiels

Un réseau immunitaire artificiel (Artificial Immune Network AIN) est un modèle de calcul de la famille des SIA qui utilise les idées et les concepts de la théorie du réseau immunitaire, principalement, les interactions entre les anticorps (stimulation et suppression), et les processus de clonage et mutation. Le déroulement d'un algorithme de réseau immunitaire peut être résumé comme suit [Tim13] :

1. Générer aléatoirement un ensemble d'anticorps du réseau (cellules mémoires)
2. Pour chaque antigène de la base d'apprentissage :
   (a) Évaluation de la similarité avec toutes les cellules mémoires ;

(b) Clonage des cellules avec les plus grandes similarités proportionnellement à ces valeurs ;

(c) Mutation de l'ensemble des clones inversement à leurs valeurs de similarités ;

(d) Ajout des meilleurs clones mutés à l'ensemble des cellules mémoires ;

(e) Évaluation des cellules mémoires et suppression de celles avec une valeur de similarité (avec l'antigène) inférieure à un seuil prédéfini ;

(f) Évaluation des cellules mémoires entres elles et suppression des paires qui ont des similarités inférieurs au seuil ;

3. Evaluation des cellules mémoires restantes entre elles et suppressions des paires qui ont des similarités inférieurs au seuil du réseau (défini par l'utilisateur) ;
4. Introduction de nouvelles cellules aléatoires à l'ensemble des cellules mémoires ;
5. Répéter les étapes 2,3 et 4 jusqu'à ce qu'un critère d'arrêt soit atteint (nombre maximum de générations).

2.5.2.2 Étude bibliographique

Basés sur la théorie du réseau immunitaire proposé par Jerne ([Jer74]), de nombreux chercheurs ont développé des modèles qui utilisent les idées et les concepts de cette théorie, pour la résolution de problèmes dans différents domaines tels que l'analyse des données, la reconnaissance des formes, la navigation autonome et l'optimisation de fonctions. Dans ce cadre Timmis et al. [TNH00] ont proposé un réseau immunitaire artificiel nommé AINE (Artificial Immune NEtwork) pour effectuer l'analyse des données. L'algorithme AINE emploie des cellules nommées Artificial Recognition Ball (ARB) pour représenter les anticorps identiques. Un lien entre deux anticorps est créé si la similarité (distance) entre deux ARB est inférieure au seuil de similarité du réseau. Les auteurs de [TN01] ont développé RLAIS, un système immunitaire artificiel de ressources limitées (Resource Limited Artificial Immune System) basé sur AINE. Les principales améliorations dans leur modèle sont un nombre fixe de cellules, présentées dans la population des ARBs avec un contrôle centralisé, chaque ARB est en compétition pour allouer des ressources. Les ARB sans ressources sont éliminées du réseau. Les processus de clonage et de mutation, ainsi que les interactions entre les anticorps sont effectuées au niveau des ARBs. C'est un mécanisme efficace de contrôle de la population qui empêche la croissance exponentielle de la taille du réseau, comme cela a été répandu dans le SIA original. Dans [N$^+$02], une approche également basée sur AINE nommé SSAIS (Self-Stabilizing Artificial Immune System) été présenté. La différence la plus importante de RLAIS est qu'il n'y a pas de quantité fixe de ressources à distribuer, et le contrôle est décentralisée au niveau des ARBs. Un algorithme sous le nom de aiNet a été proposé dans [dCVZ01]. Le réseau de cellules est généré en fonction de la distance euclidienne, il partage certaines caractéristiques de AINE, mais il en diffère en ce que la structure du réseau immunitaire ne fait pas partie du processus de clonage et de sélection d'anticorps. Comme nous l'avons déjà dit, aiNet est une simple extension de CLONALG, mais exploite les interactions entre les cellules mémoires selon la théorie du réseau immunitaire. La principale différence entre les deux méthodes est que, après l'intégration des nouveaux clones à la population de cellules mémoires, une fonction

de suppression du réseau est employée pour éliminer les cellules qui ont des valeurs de similarités inférieures au seuil du réseau. aiNet a initialement été conçu pour l'analyse de données, mais au fil des années, il a été étendu en tant qu'un outil de classification hiérarchique dans [dCT02b], puis comme modèle d'optimisation de fonctions multimodales (opt-aiNet) dans [dCT02a]. Un système immunitaire artificiel pour la classification de courrier électronique AISEC (Artificial Immune System for E-mail Classification) a été développé dans [SFT03]. L'algorithme permet l'apprentissage continu en vue de la classification en deux classes, et est utilisé pour la tâche de tri des e-mails intéressants et sans intérêt. Les auteurs de [Lv07] ont proposé un algorithme d'optimisation de fonctions multimodales. Les problèmes tels que les phénomènes de convergence prématurée et la précision insatisfaisante ont été surmontés par l'utilisation d'un réseau immunitaire chaotique. Un critère d'arrêt et utilisé pour la convergence prématurée, et la précision est améliorée grâce à une variable du chaos. Dans [HJ08], les auteurs ont proposé un réseau dit réseau de clustering (regroupement) de noyau immunitaire IKCN (Immune Kernel Clustering Network) pour la segmentation non supervisée d'images. Le travail combine les réseaux immunitaires artificiels et les séparateurs à vase marge (SVM : [HDO$^+$98]). Dans ce modèle, les caractéristiques de l'image sont divisées en sous-ensembles par les cellules du réseau, chaque sous-ensemble est tracé dans une hypersphère dans un espace de caractéristique de grande dimension par un noyau Mercer. Et finalement, un arbre couvrant minimal est utilisé pour déterminer automatiquement le nombre final de clusters. Certaines des approches de réseaux immunitaires ont été précédemment développées comme une extension pour les méthodes basées sur la sélection-clonales. Toutefois, étant donné que l'analogie a été plus explorée tout au long des dernières années, des approches novatrices intéressantes ont également été publiées dans la littérature, afin de fournir une solution d'autres problèmes, tels que la robotique et l'exploration de données. Récemment, on trouve des applications intéressantes de la théorie des réseaux immunitaires dans des systèmes automatisés, comme dans [KDS12], le réseau immunitaire est appliqué à un système de coopération multi-robots autorégulé, le robot est modélisé sous la forme d'un anticorps et son environnement d'interaction est modélisé en tant qu'un antigène. L'objectif de l'application est d'assurer la coordination et la coopération entre robots. D'autres applications basées sur les réseaux immunitaires artificiels sont présentées dans [SD16]. Ces applications concernent essentiellement la robotique, l'extraction de caractéristiques, la reconnaissance de la parole, etc.

2.5.3 Sélection Clonale

Si les immunologistes étaient invités à nommer un seul événement qui marque le début de la révolution immunologique, la plupart voterait pour l'apparition de «La théorie sélection clonale de l'immunité acquise» par Macfarlane Burnet en 1959 [B$^+$59]. Cette théorie a fourni, pour la première fois, un cadre conceptuel basé sur la biologie pour le développement des réponses immunitaires, et ses arguments principaux sont restés valables à ce jour. Il est donc, à juste titre, devenu l'alphabet de la pensée immunologique, et il est maintenant "dans le sang" de chaque immunologiste [Nag14].

Selon Burnet, la sélection clonale biologique décrit les principes basiques d'une réponse immunitaire adaptative à une attaque antigénique [B$^+$59] (voir section 2.3.3). La

sélection clonale fonctionne sur les lymphocytes T et B. Dans le cas des cellules B, alors que les récepteurs antigéniques se lient à un antigène, la cellule B est activée et commence à se multiplier en produisant de nouveaux clones qui sont une copie exacte de la cellule mère. Les clones sont ensuite soumis à une mutation et produisent des anticorps spécifiques à l'antigène envahissant [BZ93]. Après la prolifération, les cellules B se différencient en cellules plasmatiques ou des cellules mémoires de longue durée de vie. Les cellules plasmatiques produisent de grandes quantités d'anticorps qui se fixent à l'antigène en vu de l'éliminer, et les cellules mémoires aident le système immunitaire à avoir un effet protecteur sur de longues périodes de temps.

2.5.3.1 Sélection Clonale Artificielle

En plus d'être essentiellement correcte, la théorie de Burnet offre plusieurs avantages. Tout d'abord, elle permet à l'organisme de diriger plusieurs réponses immunitaires (différentes) simultanément, un avantage important dans un monde en proie aux pathogènes. En outre, l'expansion clonale pour la production continue d'anticorps après élimination de l'antigène, ainsi que la réponse secondaire améliorée après un premier passage de l'antigène.

Les chercheurs ont tenté de tirer une certaine inspiration du processus de sélection clonale, en particulier, le processus de maturation de l'affinité des cellules B avec un antigène, ainsi que le mécanisme de mutation. L'idée établie est que seules les cellules correspondantes à un antigène sont sélectionnées pour répondre à ce dernier. Une forte correspondance (similarité) entraîne un clonage important de la cellule B, tandis-qu'une faible correspondance conduit à un clonage réduit. Ces clones sont mutés à un taux inversement proportionnel à leur mesure de correspondance : plus l'anticorps correspond à l'antigène moins il sera muté, et vice-versa. En utilisant ces deux caractéristiques, l'algorithme de sélection clonale plus populaire, CLONALG a vu le jour [DCVZ02].

2.5.3.2 Étude bibliographique

La forme artificielle de la sélection clonale a été principalement popularisée par de Castro et Von Zuben, en commençant par un algorithme qu'ils ont appelé CSA (Clonal Selection Algorithm) [DCVZ00], qui a ensuite été modifié et renommé CLONALG (CLONal ALGorithm) [DCVZ02]. CLONALG génère une population de cellules mémoires (anticorps), spécifiant chacune une solution aléatoire au problème. A chaque itération certaines des meilleurs cellules mémoires existantes sont sélectionnées, clonées et mutées afin de construire une nouvelle population de cellules candidates. Les nouvelles cellules sont ensuite évaluées et un pourcentage des meilleurs de ces cellules candidates sont ajoutées à la population des cellules mémoires d'origine. Enfin, un pourcentage des pires cellules des générations précédentes est remplacé par de nouvelles créées aléatoirement.

CLONALG existe actuellement sous deux formes, une pour les tâches d'optimisation et l'autre pour la reconnaissance de formes. Cependant, l'algorithme de reconnaissance de formes a été pleinement examiné, une étude approfondie et comparative de certaines approches portant sur l'amélioration de CLONALG, avec les nouvelles recherches sur ces algorithmes sont discutées dans le chapitre suivant (section 3.2.1). D'autres algorithmes

similaires à CLONALG existent dans la littérature, par exemple [KT03] et [CNP04].

La théorie de sélection clonale a contribué dans plusieurs aspects au le développement de différents systèmes. La mémoire immunitaire et le clonage ont été explorés, et les processus de mutation des clones, ainsi que la sélection en elle même ont été exploitées offrant à la branche de sélection clonale artificielle plusieurs algorithmes, depuis CLONALG jusqu'à quelques algorithmes à objectifs multiples et bien définis, afin de résoudre des optimisations plus complexes, ou des problèmes d'apprentissage et de classification.

Dans [RHL03], les auteurs ont proposé un algorithme adaptatif de sélection clonale nommé AICSA (Adaptive Immune Clonal Strategy Algorithm). La sélection clonale combine la recherche globale et la recherche locale, de sorte que la tâche d'apprentissage peut être facilement effectuée en utilisant AICSA. Un algorithme de sélection clonale amélioré basé sur CLONALG a été proposé dans[YH04]. Un opérateur d'apprentissage a été introduit à l'algorithme de sélection clonale pour accroître la mécanisme d'apprentissage de CLONALG, et pour améliorer l'efficacité de détection. Les auteurs de [CGIR05] ont proposé un algorithme de sélection clonale avec codage réel (Real-Coded Clonal Selection Algorithm RCSA) pour l'optimisation de la conception électromagnétique. Certaines modifications ont été apportées à l'algorithme de sélection clonale pour permettre le traitement des variables à valeurs réelles pour les problèmes d'optimisation. RCSA a quelques paramètres à ajuster tels que le nombre de clones, l'intervalle de mutation, la proportion de la population sélectionnée à chaque génération, et le nombre de générations. AIRS [WTB04] est un algorithme de sélection clonale qui a été développé à partir du réseau immunitaire AINE. AIRS utilise de nombreux éléments de son patrimoine réseau immunitaire, comme la limitation des ressources pour contrôler la taille de la population, et le concept des ARB (Antigen Recognition Ball). Un ARB est une structure de données qui représente de multiples cellules mémoires identiques. L'ARB introduit également le concept des concentrations de cellules dans la sélection clonale, quelque chose qui est totalement absent dans le travail de Castro et Von Zuben, par exemple (CLONALG). Diverses améliorations de l'algorithme AIRS ont été publiées dans [FZ12], [Wat05] et récemment dans[WCA14](voir section 3.2). Un algorithme immunitaire pour la résolution des problèmes d'optimisation globale continue nommé OPT-IA (OPTimization Immune Algorithm) a été mis au point dans [CNNP05]. Trois opérateurs sont utilisés, y compris un opérateur de clonage qui fournit une population intermédiaire, et l'opérateur de mutation, où le nombre de mutations d'un clone est inversement proportionnel à sa valeur de similarité. En outre, l'opérateur de vieillissement est utilisé pour éliminer, des populations actuelles, les solutions candidates les plus anciennes, afin d'introduire la diversité et éviter les minima locaux durant le processus de recherche. OPT-IA a été amélioré dans [CNP06]. Les principales modifications de cet algorithme sont l'utilisation du codage réel à la place du codage binaire, et l'introduction d'un nouvel opérateur de mutation inversement proportionnelle. Dans [CCNS08], les auteurs ont proposé la programmation immunitaire élitiste (EIP), qui introduit le concept d'élitisme, dans lequel, à chaque étape de la génération g, la meilleure solution trouvée ne peut pas être éliminé de la population. Dans [MH09], le mécanisme d'exclusion compétitive de la sélection clonale a été discuté. Basés sur les aspects de production et de filtrage, les modèles mathématiques impliqués et certains aspects biologiques sont présentés, tels que les facteurs de concurrence entre les clones. Une approche de sélection clonale à base d'agents a été publiée dans [PISM13].

La communication et la performance des agents sont définies à l'aide d'un diagrammes UML (Unified Modeling Language).

De nombreux autres algorithmes basés sur la sélection clonale ont été publiés dans la littérature, tels que dans [KPST08], [GJZ10][PL15], etc. Les principales améliorations concernent les problème de classification et d'optimisation multi-objective ainsi que la sécurité des réseaux internet.

2.5.4 Bref résumé sur les approches SIA hybrides

Le développement de systèmes intelligents à travers l'hybridation est l'une des tendances à venir. Ces techniques sont la combinaison de différentes approches d'apprentissage automatique telles que la logique floue, le calculs évolutionnaires, les réseaux de neurones artificiels (RNA), les algorithmes génétiques (AG) et les systèmes immunitaires artificiels. Le motif principal est de surmonter les limites de l'algorithme individuel, en combinant différentes techniques d'apprentissage et d'adaptation, par lesquels il est possible d'obtenir des effets de collaboration. Au cours des dernières années, les approches intelligentes sont utilisées pour le développement d'un grand nombre de nouveaux systèmes intelligents.

Dans ce contexte, les scientifiques ont proposé des systèmes hybrides qui combinent les SIA avec d'autres techniques de calcul intelligent. Les premiers travaux qui combinent les SIA avec des algorithmes génétiques ont été développés dans [HYL97], les auteurs ont utilisé des réseaux immunitaires pour améliorer la convergence des algorithmes génétiques. Un modèle SIA flou a été proposé dans [NUCG03], l'algorithme utilise un ensemble flou pour modéliser la zone d'influence de chaque cellule mémoire, ce qui le rend plus robuste au bruit. Les auteurs de [VdCMVZ03] ont conçu un système combiné nommé CLARINETTE pour la navigation autonome, en associant les points forts des systèmes de classification, les algorithmes évolutionnaires, et un modèle de réseau immunitaire. Une méthode de classification floue non supervisée basée sur l'algorithme de sélection clonale a été publiée dans [XLT05] pour la détection d'anomalies. L'approche proposée surmonte deux inconvénients de l'algorithme k-means classique, qui est sensible à l'initialisation et tombe facilement dans les optimum locaux. Dans [FLT07], les auteurs ont présenté un réseau immunitaire artificiel hybride qui utilise l'apprentissage par essaim (swarm learning) pour accélérer la convergence du système immunitaire artificiel. Un algorithme immunitaire artificiel multi-objectifs a été proposé dans [AKA11], pour optimiser le noyau et les paramètres du SVM (Séparateur à Vaste Marge). Dans l'étape d'apprentissage du SVM, plusieurs solutions sont trouvées en utilisant l'algorithme immunitaire artificiel, ces paramètres sont évalués dans l'étape de test. L'algorithme est appliqué au diagnostic de défauts des moteurs à induction et aux problèmes de détection d'anomalies.

Il existe de nombreuses possibilités d'approches hybrides, car il y a beaucoup de paradigmes à considérer. Les approches citées ci-dessus sont quelques exemples de la façon dont les SIA peuvent améliorer ou être améliorés par d'autres techniques pour la résolution de problèmes difficiles, quant à leur complexité (SIA), un seul algorithme peut ne pas être suffisant.

2.6 Conclusion

Ce chapitre introduit le système immunitaire naturel (SIN) et discute notamment les rôles des divers organes et cellules immunitaires. Il aborde également les principes et les mécanismes importants du SIN pour décrire le comportement et les processus immunologiques lors d'une réponse immunitaire.

Inspirés par les différentes caractéristiques du SIN, trois mécanismes immunitaires ont principalement été utilisés pour le développement de différents modèles et algorithmes des Systèmes Immunitaires Artificiels (SIA). Ceux-ci comprennent la théorie des réseaux immunitaires, les mécanismes de la sélection négative et les principes de la sélection clonale.

Le chapitre présente et discute brièvement chacun de ces modèles et algorithmes, ainsi que leurs applications dans les problèmes du monde réel. En outre, les systèmes intelligents hybrides combinant les systèmes immunitaires artificiels avec d'autres techniques de calcul intelligent ont été mis en évidence.

Les avantages de la sélection clonale artificielle présentés précédemment, nous ont conduits à l'exploiter pour la classification du cancer du sein. Nous présenterons dans les chapitres suivants trois applications basées sur le principe sélection clonale. Les deux premières visent la version de reconnaissance des formes, dont la première présente différentes améliorations de l'algorithme CLONALG, et la seconde porte sur l'amélioration des taux de calcul. La dernière application est consacrée à un problème d'optimisation multimodale basé sur le principe de sélection clonale. Les chapitres 3, 4 et 5 détaillent respectivement chacune de ces applications.

Chapitre 3

CONTRIBUTION À L'AMÉLIORATION DE LA SÉLECTION CLONALE

3.1 Introduction

Comme nous l'avons précisé dans le chapitre précédent, on s'intéresse dans cette thèse spécifiquement aux algorithmes de Sélection Clonale. Dans ce chapitre, nous proposons des améliorations à un algorithme de sélection clonale artificielle : l'algorithme CLONALG. En effet, cet algorithme est très utilisé et largement référencé dans la littérature. Son principe consiste à simuler d'une manière simple et efficace le comportement de la sélection clonale biologique. Il est basé sur un cycle répété de sélection (ou re-sélection), clonage, mutation et génération de nouvelles cellules pour maintenir la diversité dans l'algorithme.

Dans un premier temps, on abordera les différents détails et principes des algorithmes de sélection clonale, ainsi que les travaux récents du domaine. Ensuite, on se focalisera sur l'algorithme CLONALG pour traiter les observations faites sur ce dernier. Ces observations concernent les étapes d'initialisation et maintenance de la diversité, qui sont faites de manières aléatoires. En effet, la diversité d'une cellule mémoire réside dans sa capacité de reconnaître plusieurs types d'antigènes.

Afin de renforcer l'apprentissage de CLONALG avec une meilleure initialisation et une diversité maîtrisée, trois différentes méthodes sont proposées dans ce chapitre appelées Median Filter Clonal ALGorithm (MF-CLONALG), Average Cells Clonal ALGorithm (AC-CLONALG) et Validity Interval Clonal Selection (VI-CS). L'évaluation de ces approches est faite sur les bases de données WDBC et DDSM présentées dans les annexes A et B.

3.2 Développement de la Sélection Clonale Artificielle

Depuis 1959, il y a eu des améliorations de la théorie de Burnet [B$^+$59], par rapport à la façon dont les antigènes sont reconnus. Mais les principes de base de la sélection clonale et la maturation de la similarité par hypermutation sont suffisantes aux fins des algorithmes de sélection clonale artificielle.

Dans cette section, nous discuterons d'abord du développement des algorithmes de sélection clonale artificielle, puis nous nous concentrerons sur l'algorithme CLONALG. Nous présenterons les différentes étapes et aborderons quelques observations faites sur ce dernier.

Tous les algorithmes à base de sélection clonale artificielle (SCA) présentés dans le chapitre précédent se concentrent essentiellement autour d'un cycle répété de sélection, clonage, mutation et remplacement (génération de nouvelles cellules pour maintenir la diversité dans l'algorithme). De nombreux paramètres peuvent être accordés, y compris le taux de clonage, le nombre initial d'anticorps et le taux de mutation pour les clones. CSA [DCVZ00], CLONALG [DCVZ02] et AIRS [WTB04] intègrent tous cette fonctionnalité de base.

Artificial Immune Recognition System (AIRS) est une forme d'algorithme de sélection clonale qui a été développé à partir du réseau immunitaire AINE (Artificial Immune NEtwork [TNH00]). Cet algorithme a été proposé par A. Watkins en 2001 dans sa thèse de Master à l'université de Mississipi [Wat01]. L'amélioration de AIRS a été faite en 2004 dans [WTB04], les auteurs optimisent le temps d'exécution et le nombre de cellules mémoires générées. Cette méthode est dotée par la possibilité d'apprentissage parallèle prouvé dans le PHD de A. Watkins à l'université de Kent en 2005 [Wat05]. Le diagramme général de AIRS est donné dans la figure 3.1.

AIRS utilise de nombreux éléments de son patrimoine de réseau immunitaire, comme la forme inhabituelle de limitation de ressources pour contrôler la taille de population ainsi que le concept des *ARBs* (Antigen Recognition Ball). Divers travaux proposant des améliorations sur l'algorithme AIRS ont été publiés, notamment sur l'évolution des cellules mémoires, le seuil de similarité, etc. Le premier travail visant l'amélioration de AIRS a été proposé par A.Watkins et J.Timmis en 2004, dans [WT04a]. Certaines complications inutiles dans l'algorithme original on été corrigées. En effet, dans la version initiale d'AIRS, le développement des cellules mémoires d'une classe donnée est effectué en utilisant les populations des ARBs de toutes les classes. La sélection des meilleurs cellules est basée sur leurs valeurs de similarité et leurs appartenances, c-à-d, les cellules sélectionnés pour le clonage sont soit celles avec les plus grandes similarités avec l'antigène s'ils appartiennent à la même classe, ou-bien, celles des autres classes avec les plus faibles valeurs de similarité. Alors que normalement, le développement de la cellule candidate doit se faire uniquement avec les ARBs de la même classe. Les auteurs ont aussi observé une certaine redondance dans les cellules mémoires produites, et que ces dernières ne semblent pas d'une bonne qualité.

Les améliorations apportées dans ce travail consistent à utiliser seulement les ARBs de la même classe pour l'évolution de la cellule mémoire candidate. A la fin de l'apprentissage

de chaque antigène, la population d'ARBs est vidée. La seconde amélioration a pour but d'explorer le rôle de la mutation dans la qualité des cellules mémoires générées. Le processus de mutation a été modifié de sorte que la quantité de mutation d'une cellule est contrôlée par sa valeur de stimulation, c.à.d. plus la valeur de stimulation est grande, moins l'intervalle de mutation est large.

Récemment, on trouve d'autres versions de l'algorithme AIRS dans divers domaines d'application. Dans [FZ12] , les auteurs proposent une version améliorée de AIRS nommée IARS (Improved Artificial Recognition System). L'amélioration consiste à remplacer le seuil de similarité dans AIRS qui est la mesure d'affinité moyenne entre tous les exemples d'apprentissage de toutes les classes, par un seuil de similarité spécifique à chaque classe à apprendre. Cela permet de générer une meilleure population de cellules mémoires améliorant le taux de classification et réduisant les taux de calcul.

Une autre version améliorée de AIRS a été publiée dans [WCA14]. Les auteurs proposent une approche hybride de l'algorithme AIRS et l'algorithme de test de signe opposé. La méthode à été validée par l'application sur plus de 40 différentes bases de données répertoriées dans UCI et Keel.

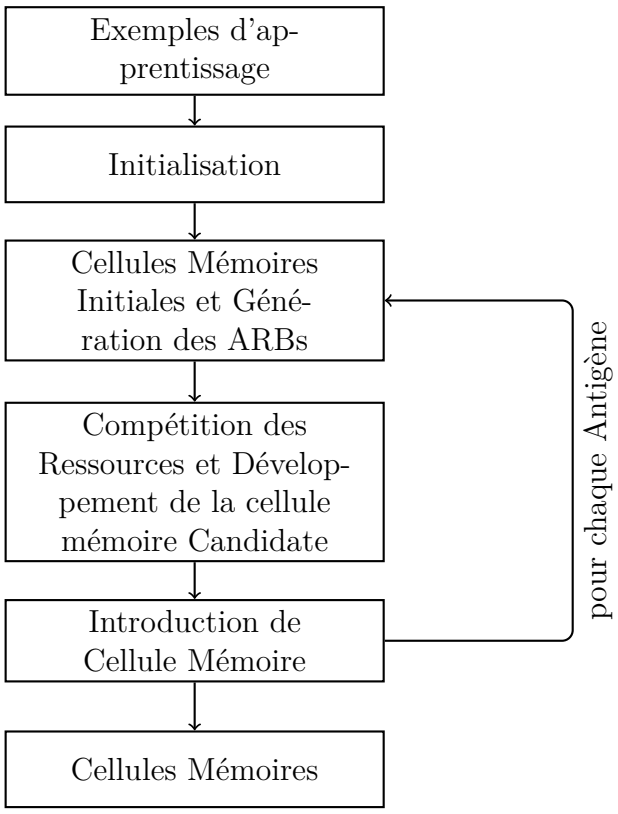


FIGURE 3.1 – Diagramme d'apprentissage de AIRS

Un autre algorithme nommé CLONALG, basé sur le principe de sélection clonale a largement été utilisé pour résoudre divers problèmes complexes d'ingénierie tels que la reconnaissance des formes, l'optimisation combinatoire, la détection d'anomalies. etc. En raison de sa simplicité à simuler efficacement le comportement de la sélection clonale, il

y a eu davantage l'accent sur cet algorithme.

L'Algorithme de sélection clonale (CSA) a d'abord été proposé par De Castro et Von Zuben dans [DCVZ00], il a ensuite été amélioré et nommé CLONALG dans [DCVZ02]. C'est l'un des algorithmes les plus réputés dans le domaine des Systèmes Immunitaires Artificiels qui a initialement été proposé pour effectuer la reconnaissance des formes, et qui a été adapté pour effectuer des tâches d'optimisation.

3.2.1 Sélection clonale artificielle par CLONALG

CLONALG suit la théorie de base d'une réponse du système immunitaire aux agents pathogènes. En gros, les composants de l'algorithme sont des cellules nommées cellules mémoires, et un antigène qui est un envahisseur qui attaque le système immunitaire. Le principe de cet algorithme consiste à construire aléatoirement une population de cellules mémoires initiales et de les exposer aux antigènes (exemples d'apprentissage) d'une manière répétitive, pour un nombre défini de générations. Le but étant de développer une population de cellules mémoires plus spécifique à ces antigènes, grâce aux processus de clonage et de mutation. Les étapes principales de l'algorithme CLONALG sont :

- Génération des cellules mémoires initiales par une sélection aléatoire des antigènes (exemples d'apprentissage).
- Évaluation des cellules mémoires et sélection des plus représentatives de l'antigène en cours d'apprentissage.
- Clonage, mutation et re-sélection du meilleur clone muté.
- Maintenance de la diversité par le rejet des cellules mémoires moins représentatives de l'antigène, et leur remplacement par d'autres aléatoirement générées.

Comparé à d'autres algorithmes de sélection clonale tels que AIRS, CLONALG présente une faible complexité et nécessite moins de paramètres qui peuvent influencer la précision de classification [Zha11] [Bro05]. Il a été appliqué avec succès pour résoudre différents problèmes complexes, et offre une précision prometteuse dans le domaine de reconnaissance des formes.

La première amélioration de CLONALG a été faite en 2003 dans [WG03]. Le but de ce travail est de valider l'algorithme et établir sa performance pour la reconnaissance de formes. Certaines fonctionnalités supplémentaires ont été mises en œuvre. Les auteurs ont remarqué une caractéristique potentiellement négative dans CLONALG qui est son incapacité à capitaliser sur les informations générées par la population des clones. En effet, une fois la cellule mémoire sélectionnée, les clones mutés restant sont éliminés même si la population des clones peut contenir un certain nombre de cellules candidates de hautes similarités. Alors qu'en préservant une plus grande proportion de la population mûrie, l'algorithme pourrait construire une population de cellules mémoires avec des correspondances de plus hautes affinités en moins de générations. Cependant, cette solution introduit le risque de se coincer sur un minimum local. Pour éviter que la population devient de plus en plus étroite, les auteurs ont introduit une certaine modification à la phase de remplacement aléatoire des cellules mémoires de faibles affinités, pour maintenir

la diversité et converger vers une solution optimale. Ils proposent un CLONALG amélioré nommé CLONCLAS, avec une étape supplémentaire qui consiste à remplacer les cellules mémoires avec les meilleurs clones mutés, avant de rejeter les plus faibles cellules et de les remplacer par des cellules aléatoires.

A.Sharma et D.Sharma ont proposé dans [SS11] une variante de CLONALG dédiée aux tâches de classification appelée CLONAX (Clonal Selection Algorithm for Classification). L'algorithme produit des cellules mémoires généralisées qui reconnaissent mieux les antigènes de la même classe. Même si les cellules mémoires initiales sont sélectionnées aléatoirement comme dans CLONALG, les auteurs proposent de définir au préalable le nombre de ces cellules. Pendant l'Apprentissage, au lieu de sélectionner un seul clone muté, CLONAX crée des cellules mémoires généralisées en sélectionnant un nombre k de clones mutés ayant les plus grandes similarités avec l'antigène à apprendre. La similarité moyenne de chacune de ces k cellules est calculée par la suite, et un filtrage des meilleurs clones est effectué pour réduire la possibilité de bruit dans les données. Le filtrage consiste à supprimer un clone si sa similarité avec au moins deux antigènes d'une classe différente est supérieure à sa valeur de similarité moyenne. Par contre, si la similarité moyenne d'un clone est inférieure à celle avec un seul antigène d'une autre classe, l'antigène est considéré comme bruit et supprimé de la base d'apprentissage.

La création des cellules mémoires généralisées pour chaque classe est une idée intéressante pour mieux représenter les données à apprendre. Par contre, il y a une contradiction dans le fait de supprimer des exemples des classes d'apprentissage afin d'améliorer leur représentativité. En effet, la suppression de ces exemples peut conduire à une perte d'informations, ce qui aura une influence sur les résultats de l'algorithme. A la fin de l'apprentissage de chaque antigène, et afin de maintenir la diversité dans l'algorithme, CLONAX rejette les cellules mémoires avec les plus faibles valeurs de similarités, et les remplace par d'autres générées aléatoirement exactement comme le fait CLONALG. Le remplacement de ces cellules est effectué sans vérifier si les cellules aléatoires sont plus représentatives que celles qui sont rejetées.

Ce problème de rejet des cellules mémoires pour la diversité a été abordé dans [TRR14]. En effet, dans cette version améliorée de CLONALG, les cellules mémoires avec une faible similarité sont d'abord remplacées par les meilleurs clones mutés. Ensuite, les cellules mémoires restantes sont remplacées par d'autres aléatoirement générées seulement si elles ont une plus grande affinité. Par contre, la vérification est faite par rapport à l'antigène en cours d'apprentissage seulement, alors que la diversité d'une cellule mémoire réside dans sa capacité de reconnaître plusieurs antigènes de la même classe.

Les différents travaux cités ci-dessus proposent d'utiliser les informations contenues dans les clones mutés au lieu de les rejeter directement après la re-sélection de la cellule candidate. CLONCLAS [WG03] remplace les cellules mémoires de faibles affinités par les meilleurs clones mutés, tandis-que CLONAX [SS11] sélectionne plusieurs clones pour rejoindre la population des cellules mémoires, et procède à un filtrage par la suite. Mais à notre connaissance, aucun travail ne s'est intéressé à la pertinence des cellules mémoires

aléatoires ajoutées à la fin d'apprentissage de chaque antigène ; ni aux cellules mémoires initialement sélectionnées pour lancer l'apprentissage de l'algorithme.

En effet, l'étape d'initialisation de CLONALG consiste à sélectionner aléatoirement quelques exemples de chaque classe d'apprentissage, pour constituer l'ensemble des cellules mémoires initiales. Mais il n'y a aucune certitude que ces exemples représentent bien toutes les données à apprendre. Il y a effectivement des bases de données où les écarts entre les attributs des exemples de la même classe peuvent être grands, et la sélection de quelques exemples seulement pour lancer l'apprentissage peut influencer les résultats de l'algorithme. Il est donc important de trouver une méthode pour générer des cellules mémoires initiales qui représentent toutes les données à apprendre. Pour remédier au problème d'initialisation dans l'algorithme CLONALG, nous proposons une solution d'amélioration dans la section 3.3.

En effet, à la fin d'apprentissage de chaque exemple (antigène), les cellules mémoires ayant les plus faibles similarités avec ce dernier sont supprimées et remplacées par d'autres aléatoirement générées. Tous les travaux améliorant CLONALG continuent d'utiliser cette méthode pour maintenir la diversité, et éviter les minima locaux. Mais aucune de ces méthodes ne vérifie la pertinence des cellules rejetées par rapports aux autres exemples de la classe, ni celle des cellules aléatoires ajoutées. Le travail dans [TRR14] propose de vérifier d'abord que les cellules rejetées sont moins bonnes que celles qui vont les remplacer. Mais l'évaluation est faite uniquement par rapport à l'antigène en cours d'apprentissage. Alors que les cellules rejetées peuvent être les plus représentatives des autres exemples de la base d'apprentissage, ou bien produire des cellules mémoires de haute qualité dans les générations qui suivent. La figure 3.2 fournit un organigramme de l'algorithme CLONALG mentionnant les limitations observées. Les différentes améliorations effectuées sur cet algorithme sont détaillées dans la section 3.3..

3.3 Contributions apportées à l'algorithme CLONALG

Afin d'améliorer les taux de classification du cancer du sein dans les bases référencées, nous avons proposé des contributions permettant l'amélioration de l'algorithme CLONALG. Dans ce cadre, et pour traiter les limitations mentionnées dans les sections précédentes, nous avons apporté quelques modifications à CLONALG. En effet, pour garantir une population de cellules mémoires initiales qui représente la totalité des données à apprendre, on propose dans un premier temps, de créer ces dernières à partir de sous-groupes locaux de chaque classe d'apprentissage. On s'intéresse ensuite à la diversité de CLONALG. Quelques améliorations sont introduites à l'étape d'apprentissage par la création de cellules mémoires pertinentes pour le clonage, et la généralisation de ces dernières grâce à un intervalle de validité pour une diversité maîtrisée.

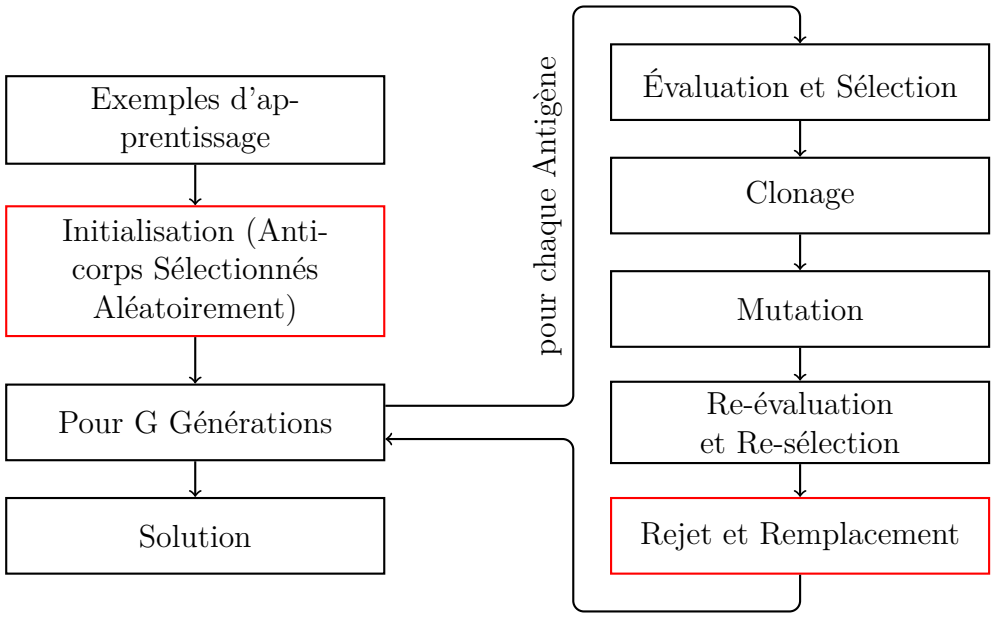


FIGURE 3.2 – Organigramme de l'algorithme CLONALG

3.3.1 Étape d'initialisation

L'étape d'initialisation est très importante avant de lancer n'importe quel algorithme d'apprentissage. Dans CLONALG, l'initialisation consiste à construire un ensemble de cellules mémoires initiales pour chaque classe à apprendre. Cela est fait en sélectionnant aléatoirement quelques exemples à partir de la base d'apprentissage. Ces cellules initiales vont contribuer à la génération de la population des cellules mémoires finales de chaque classe.

Tous les travaux améliorant CLONALG évoquent la nécessité d'explorer la population des clones pour en extraire le maximum d'informations. Mais ce qu'il ne faut pas oublier, c'est que la population des clones mutés dépend fortement des cellules mémoires initiales. Le choix aléatoire de quelques exemples d'apprentissage n'est donc pas suffisant, car ces exemples ne contiennent pas forcément toutes les informations à apprendre.

Dans ce chapitre, nous nous sommes intéressés à ce problème. Nous proposons une méthode qui généralise les cellules mémoires initiales pour représenter efficacement leurs classes d'apprentissage.

Au lieu de sélectionner aléatoirement quelques exemples de chaque classe pour constituer l'ensemble de cellules mémoires initiales, on propose de créer ces dernières d'une manière plus efficace. La création de ces cellules se fait en partitionnant les classes d'apprentissage en sous-groupes locaux de taille k par un processus aléatoire, incluant tous les exemples à apprendre. La cellule moyenne de chaque sous-groupe local est calculée par la suite. Ces cellules moyennes sont considérées comme cellules mémoires initiales créées spécifiquement pour l'apprentissage. De cette manière, on garantit que toutes les données à apprendre sont incluses dans la population de cellules initiales, ce qui permettra d'assurer un bon apprentissage de l'algorithme. La figure 3.3 décrit le processus de création des cellules mémoires initiales.

Le nombre de cellules mémoires initiales de chaque classe est relativement lié à la taille de cette dernière. A titre d'exemple, pour une classe d'apprentissage comprenant X exemples, le nombre de cellules mémoires initiales peut être choisi entre le tiers et la moitié de X.

Notons que la population de cellules mémoires initiales doit être assez consistante pour bien représenter la base d'apprentissage. Mais en même temps, le nombre de ces cellules ne doit pas être très grand, car ça influe sur la rapidité de l'algorithme.

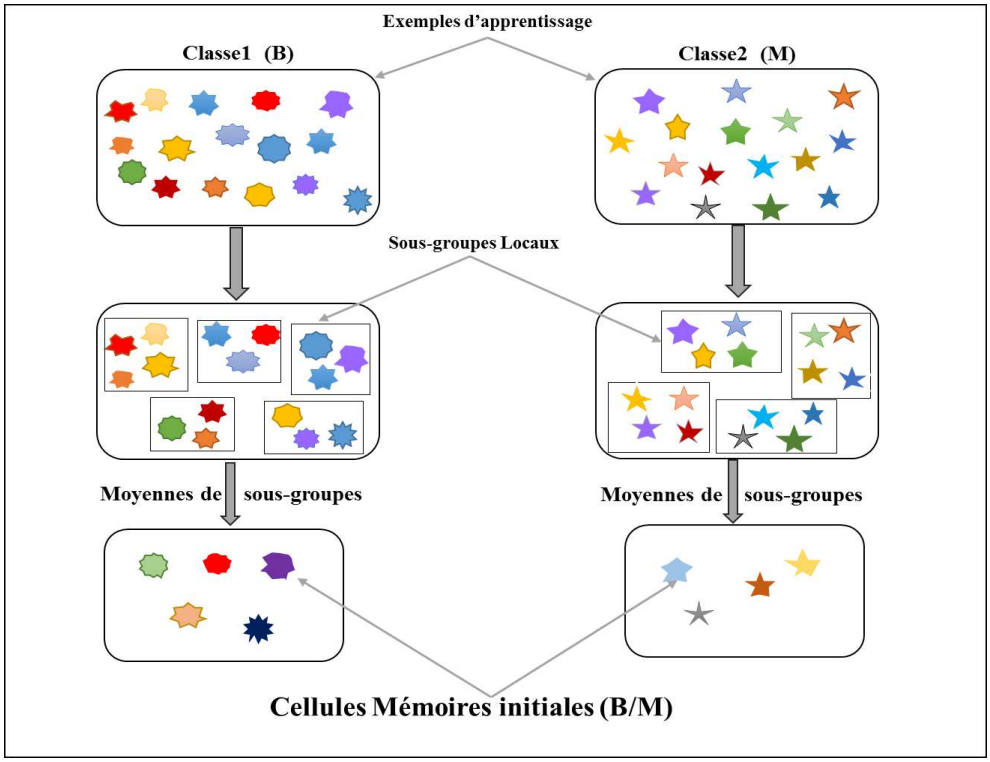


FIGURE 3.3 – Création des cellules mémoires initiales : Création des sous-groupes locaux et calcul des cellules moyennes

En utilisant cette méthode d'initialisation dans CLONALG, on garantit une meilleure représentation de la totalité des données à apprendre, ce qui implique un bon apprentissage des données.

Dans [Bro05], l'auteur mentionne que parmi les points négatifs de CLONALG l'introduction des individus aléatoires. Car, si cela permet d'échapper aux minima-locaux, les exemples aléatoires fournissent un moyen d'introduire complètement de nouveaux prototypes (exemples) qui peuvent ne pas être utiles pour le processus d'apprentissage. Pour cette raison, et afin de permettre un bon apprentissage de CLONALG avec une diversité maîtrisée tout en permettant d'éviter les minima-locaux, nous avons apporté quelques améliorations de ce dernier. Après avoir amélioré l'étape d'initialisation, la deuxième modification introduite à CLONALG a pour but d'éviter l'introduction d'individus aléatoires à l'ensemble des cellules mémoires. L'idée est de créer des cellules mémoires à partir des meilleurs individus et de les ajouter à la population des cellules mémoires si elles sont

pertinentes. Nous proposons trois approches qui améliorent la sélection dans CLONALG, et préservent la diversité de l'algorithme sans avoir a rejeter aucune cellule.

La première approche que nous avons proposé est nommée Filtre Median (Median Filter Clonal Algorithm [MF-CLONALG]). Elle consiste à créer des cellules mémoires médianes à partir des L meilleurs cellules (L étant nombre de descripteurs (attributs) de la base de données). La seconde permet la sélection clonale par cellules moyennes (Average Cells Clonal Algorithm [AC-CLONALG]). La troisième approche proposée utilise un intervalle de validité pour permettre aux cellules créées, ou les clones mutés, de rejoindre l'ensemble des cellules mémoires finales (Validity Interval Clonal Selection algorithm [VI-CS]). Le principe de chacune de ces approches est détaillé dans les sous-sections qui suivent. Les résultats expérimentaux seront discutés dans la section 3.4 .

3.3.2 Selection Clonale par Filtre Médian (MF-CLONALG)

La plupart des versions améliorant CLONALG citées dans la section précédente proposent d'explorer les informations contenues dans les clones mutés, afin de construire une population de cellules mémoires de hautes similarités. Et en vue de préserver la diversité de l'algorithme et éviter les minima-locaux, ils remplacent les cellules mémoires les plus faibles par d'autres cellules aléatoires. L'idée proposée dans ce travail consiste à explorer la population des cellules mémoires et celles des clones mutées à la fois. En effet, au lieu de sélectionner une seule cellule qui maximise la similarité avec l'antigène, on propose d'en sélectionner plusieurs. Le but est de chercher à créer des cellules mémoires pertinentes et préserver la diversité dans l'algorithme simultanément. De cette manière, on évite aussi d'introduire des cellules aléatoires qui peuvent ne pas être pertinentes, et qui constituent l'une des reproches de l'algorithme CLONALG [Bro05].

Le principe du filtre médian (MF-CLONALG) consiste à créer des cellules médianes à partir des meilleures cellules mémoires. Ces cellules médianes rejoindrons l'ensemble des cellules mémoires finales seulement si elles prouvent leur pertinence. La création d'une cellule médiane est relative à la longueur d'une cellule, c-à-d, le nombre de descripteurs de cette cellule.

Soit L le nombre de descripteurs de la base de données à apprendre. Pour chaque exemple d'apprentissage (Antigène), on sélectionne L cellules mémoires maximisant la valeur de similarité avec ce dernier. A partir de la matrice carrée M de ces L cellules (chaque ligne de la matrice constitue une cellule mémoire), on prend la valeur médiane de chacun des L descripteurs. Cela est fait en triant les colonnes de la matrice M par ordre croissant, la ligne du milieu comporte alors les valeurs médianes de chaque descripteur. La figure 3.4 présente un exemple du processus de création de la cellule médiane à partir de la base WDBC.

Après avoir créé la cellule médiane, on procède de la manière suivante :

1. Calcul de la similarité de la cellule médiane avec l'exemple d'apprentissage (AG), et comparaison avec la plus haute similarité des cellules mémoires. La cellule (mémoire ou médiane) qui présentera la plus grande similarité avec AG est sélectionnée pour l'étape suivante (on nommera cette cellule C_{Clon}).

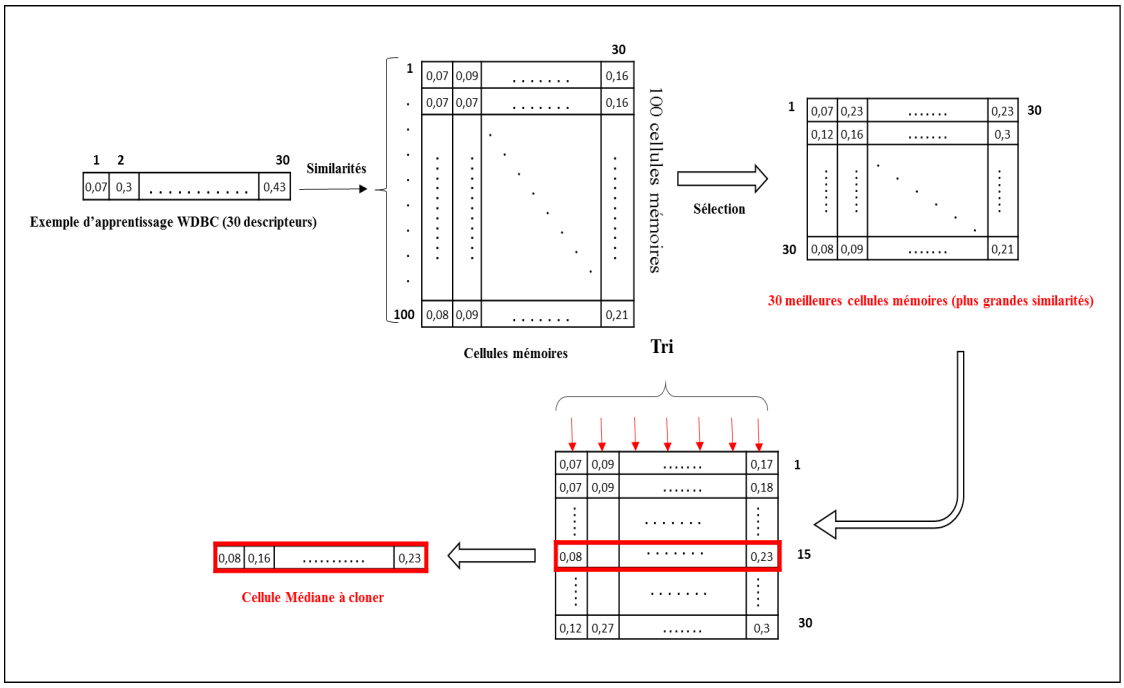


FIGURE 3.4 – Exemple de création de la cellule médiane dans MF-CLONALG (base WDBC = 30 descripteurs)

2. Clonage et mutation de la cellule C_{Clon}. Le clonage est effectué proportionnellement à la valeur de similarité de C_{Clon}, et la mutation est faite inversement à cette dernière, c-à-d, plus la similarité est grande plus on produit de clones et moins l'intervalle de mutation est large, et vice-versa.
3. Calcul des similarités des clones mutés avec l'exemple d'apprentissage AG. Seuls les clones ayant une meilleure similarité que la cellule d'origine (C_{Clon}) sont ajoutés à l'ensemble des cellules mémoires finales.

— Si C_{Clon} = cellule médiane, et aucun clone n'est pris, C_{Clon} rejoins les cellules mémoires finales.

Après avoir présenté la totalité des exemples d'apprentissage (de toutes les classes), et effectué les étapes ci-dessus pour chacun d'eux, nous dirons qu'une génération a été achevée. L'organigramme de MF-CLONALG montrant les améliorations apportées à CLONALG est donné dans la figure 3.5.

Dans cette approche proposée, l'étape d'initialisation a été améliorée grâce à la création de cellules mémoires représentatives des exemples d'apprentissage (section 3.3.1). Ensuite, dans l'étape de sélection, en vue du clonage et de mutation ; la technique de MF-CLONALG permet de construire des cellules potentielles. Ces cellules médianes contribuent à préserver la diversité de l'algorithme en évitant d'introduire des cellules aléatoires.

La seconde technique que nous avons proposé pour améliorer l'étape de sélection dans l'algorithme CLONALG est présentée dans la section 3.3.3.

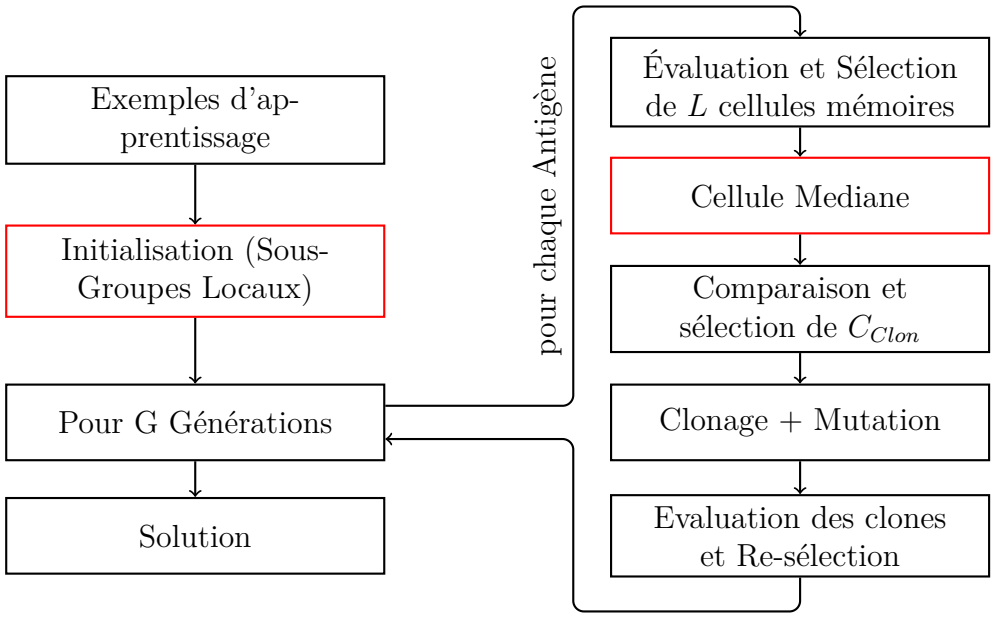


FIGURE 3.5 – Organigramme de MF-CLONALG

3.3.3 Sélection clonale par cellules moyennes (AC-CLONALG)

Comme nous l'avons précisé auparavant, même si l'algorithme CLONALG a prouvé son efficacité dans le domaine de la reconnaissance de formes, il présente néanmoins un inconvénient non négligeable, qui est l'introduction d'individus aléatoires en vue de maintenir la diversité [Bro05].

Cette section présentera la deuxième approche que nous avons proposé pour optimiser l'apprentissage de CLONALG. Le but reste le même, celui de générer une population de cellules mémoires spécifiques à chaque classe d'apprentissage, tout en évitant de rejeter des cellules potentielles à chaque présentation d'un antigène.

L'idée de AC-CLONALG s'inspire de l'amélioration que nous avons apporté à l'étape d'initialisation (section 3.3.1), et le but est de maintenir une bonne diversité dans l'algorithme. En effet, nous avons proposé dans l'étape d'initialisation de créer des cellules initiales spécifiques pour chaque classe à apprendre. A partir de ce principe, on pourrait donc créer des cellules mémoires plus pertinentes et représentatives pendant l'apprentissage aussi. Au lieu de sélectionner une seule cellule à cloner, nous procéderons à la sélection de P cellules ayant les plus grandes valeurs de similarités avec l'exemple en cours d'apprentissage (AG). Une cellule moyenne C_{moy} est calculée par la suite à partir de ces P cellules. C_{moy} est donc la cellule candidate à subir aux opérateurs de clonage et de mutation. Le but de ce processus est de maintenir la diversité dans l'algorithme en introduisant de nouvelles cellules mémoires, et contrairement à CLONALG, ces cellules sont spécifiques, pas aléatoires. De plus, ces cellules moyennes ne seront ajoutées à l'ensemble des cellules mémoires finales seulement si elles prouvent leur efficacité. Les opérations de clonage et de mutation sont appliqués sur C_{moy}, si sa valeur de similarité avec AG est plus élevée que celle de la cellule mémoire la plus proche de ce dernier. Sinon, c'est la

meilleure cellule mémoire qui sera clonée et mutée. La cellule sélectionnée pour le clonage sera nommée C_{clon}.

Après les processus de clonage et de mutation, la population des clones est explorée comme il a été suggéré dans les différentes variantes de CLONALG, mais de manière différente. Dans notre approche, contrairement aux autres travaux qui proposent d'ajouter directement plusieurs clones aux cellules mémoires finales, la pertinence de chaque clone est d'abord vérifiée. Ils rejoindrons les cellules mémoires seulement s'ils présentent une meilleure similarité que leurs cellules d'origine, c-à-d, si un clone muté est plus représentatif de AG que la cellule d'où il a été cloné (cellule mère), il est ajouté à l'ensemble des cellules mémoires. Sinon, il sera automatiquement rejeté.

Cette approche a été nommée Average Cells Clonal Algorithm (AC-CLONALG), par rapport aux cellules moyennes générées pour le clonage dans le but de produire davantage de cellules mémoires plus pertinentes. En outre, elle maintient une bonne diversité des données d'apprentissage sans pour autant rejeter d'autres cellules qui peuvent être utiles dans les générations qui suivent, ou avec d'autres exemples d'apprentissage. La figure 3.6 présente les différentes phases de l'algorithme AC-CLONAG.

Après avoir créé des cellules mémoires moyennes pour améliorer la sélection clonale, nous avons constaté que les opérations de moyennes locales ne représentent pas convenablement la diversité réelle des classes d'apprentissage. Nous avons donc pensé à calculer un intervalle de validité pour chaque classe à apprendre afin de valider les cellules mémoires et les clones générés. La section 3.3.4 présente une description détaillée de l'approche.

3.3.4 Intervalle de validité pour la sélection clonale (VI-CS)

Comme nous l'avons mentionné à la fin de la section précédente, nous avons constaté que les cellules mémoires créées, qui sont utilisées dans la classification, ne présentent pas souvent une bonne représentativité de la totalité des cellules des classes à apprendre. En effet, même si les cellules mémoires initiales ne sont pas aléatoirement sélectionnées, ces cellules déterminées localement ne garantissent pas constamment la représentativité globale de la classe.

Nous avons vu dans la section 3.2 que pour permettre une bonne représentativité des données d'apprentissage, les auteurs dans [SS11] ont proposé de générer des cellules mémoires généralisées pour chaque classe d'apprentissage. Mais le filtrage de ces cellules généralisée consiste à supprimer un antigène (exemple d'apprentissage) d'une autre classe si sa valeur de similarité avec la cellule mémoire est supérieure à la similarité moyenne de cette dernière avec les antigènes de sa classe. L'idée des cellules mémoires généralisée est très intéressante pour représenter efficacement la diversité des données dans chaque classe d'apprentissage. Par contre, il est dangereux de supprimer des exemples d'apprentissage d'une classe pour améliorer la représentativité d'une autre. Car les exemples supprimés contiennent des informations importantes sur les classes qu'ils représentent, ce qui aura une influence sur les résultats de l'algorithme.

Pour cette raison, et dans le but d'améliorer la représentativité des classes d'apprentissage, nous avons proposé dans ce travail de déterminer un intervalle de validité pour la

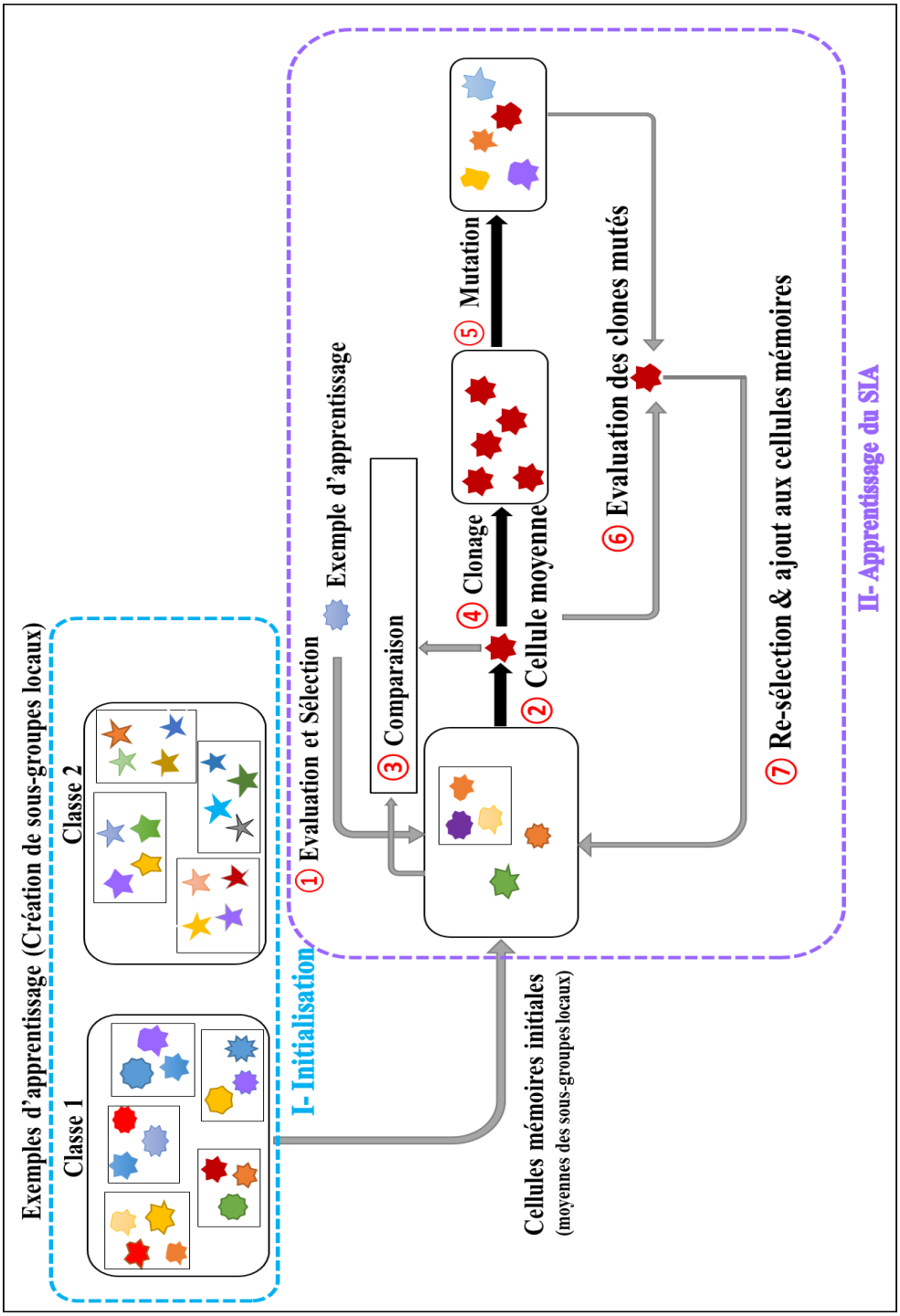


FIGURE 3.6 – Diagramme d'apprentissage de AC-CLONALG composé de : I- Etape d'initialisation : création des cellules mémoires initiales à partir de sous groupes locaux. II- Etape d'apprentissage du SIA : création des cellules mémoires initiales à partir de sous groupes locaux. 1) Évaluation des cellules mémoires et sélection de P meilleurs, 2) Création de la cellule moyenne (C_{moy}) 3) Comparaison entre C_{moy} et la meilleure cellule mémoire et sélection de C_{clon} , 4) clonage, 5) mutation des clones, 6) évaluation des clones mutés et 7) ajout des meilleurs clones mutés aux cellules mémoires finales.

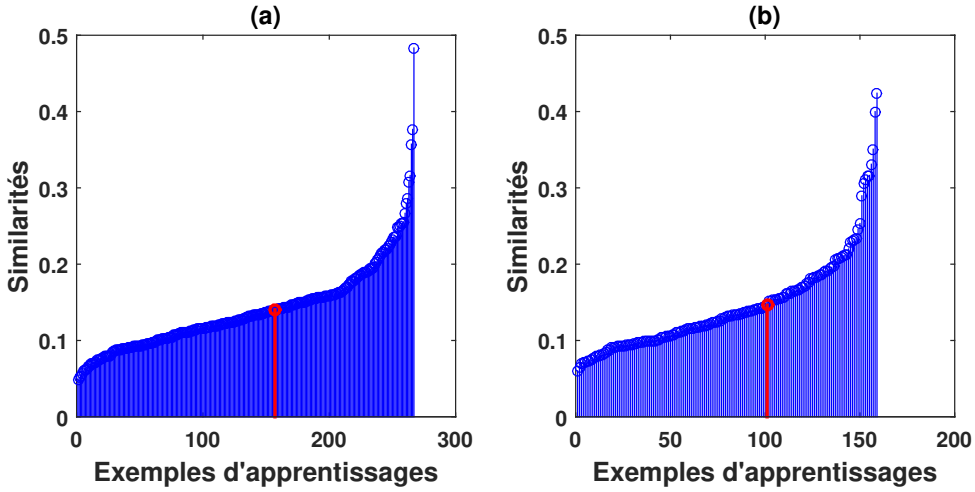


FIGURE 3.7 – Histogrammes des similarités entre la Cellule Moyenne Globale et les exemples d'apprentissage (en bleu) et similarité moyenne (en rouge) de la classe bénigne (a) et la classe maligne (b) le la base WDBC.

sélection des cellules mémoires crées plus efficacement. Ces cellules mémoires sont généralisées pour représenter la diversité des classes d'apprentissage grâce à leurs similarités moyennes.

L'approche que l'on propose est composée de trois étapes : la première consiste à calculer l'intervalle de validité de chaque classe d'apprentissage. La seconde étape concerne la sélection des cellules mémoires initiales en utilisant l'intervalle de validité. La dernière étape de l'algorithme présente l'apprentissage global du SIA.

3.3.4.1 Étape 1 : Intervalle de Validité pour la sélection (VI)

Pour chaque classe à apprendre, on calcule une Cellule Moyenne Globale (CMG) a partir des cellules d'apprentissage. On détermine par la suite la similarité de CMG avec toutes ces cellules d'apprentissage. Puis, on calcule les caractéristiques statistiques : la moyenne (Sim_moy) , la variance, et l'écart type (déviation standard).

La similarité moyenne d'une classe (Sim_moy) est la moyenne des similarités de (CMG) avec tous les exemples d'apprentissage de cette classe. La figure 3.7 montre un exemple de la base WDBC ou les histogrammes en bleu présentent les similarités entre la CMG et les exemples d'apprentissage de chaque classe ((a) : bénigne , (b) : maligne). La similarité moyenne de chaque classe est tracée en rouge.

Par la suite, on calcule l'écart-type (σ) des similarités de chaque classe par l'équation suivante :

$$\sigma = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})^2} \tag{3.1}$$

Avec $\bar{x}$ = la similarité moyenne de la classe (Sim_moy), M : le nombre total des

cellules d'apprentissage, et x_i la similarité de la cellule i avec CMG.

L'intervalle (VI) de validité de chaque classe est déterminé par :

$$VI = Sim\_moy \pm \sigma = [Sim\_moy - \sigma, Sim\_moy + \sigma] \tag{3.2}$$

Cet intervalle servira à la validation des clones sélectionnés pour rejoindre les cellules mémoires finales.

3.3.4.2 Étape 2 : Sélection des cellules mémoires initiale en utilisant VI

Dans cette étape, la création des cellules mémoires initiales pour chaque classe d'apprentissage est effectuée de la même manière détaillée dans la section 3.3.1, à savoir, par moyennes de sous-groupes locaux. L'objectif est de créer des cellules initiales représentant toutes les données à apprendre, au lieu de sélectionner aléatoirement des cellules pas suffisamment représentatives de la classe.

Aprés avoir créé les cellules mémoires initiales $(CM_1, .., CM_N)$, on calcule la similarité moyenne de chacune avec tous les exemples d'apprentissage de la même classe $(Ex_1, .., Ex_M)$.
Avec N le nombre de cellules mémoires initiales (NB_{CM_i}).

$$Sim\_moy(CM_i) = \frac{1}{M} \sum_{j=1}^{M} Sim(CM_i, Ex_j) \tag{3.3}$$

— Si $Sim\_moy(CM_i) \in VI$; CM_i est maintenue dans l'ensemble de cellules de mémoire initiales.
— Sinon, elle ne sera pas considérée.

— Si le nombre de cellules mémoires initiales restantes n'est pas consistant, on procède à une création de nouvelles cellules moyennes, et recommence le processus de sélection par VI.

Le processus de création de l'intervalle de validité (étape 1) et la sélection des cellules mémoires initiales (étape 2) sont illustrés dans la Figure 3.8.

3.3.4.3 Étape 3 : apprentissage du SIA

L'apprentissage du système immunitaire artificiel (SIA) commence à cette étape. Les cellules moyennes sont créées et ajoutées à l'ensemble des cellules mémoires finales comme nous l'avons décrit dans l'approche précédente (AC-CLONAG, Section 3.3.3), mais avec une condition supplémentaire. En effet, Les cellules générées (moyennes ou clones mutés) seront ajoutées à l'ensemble de cellules mémoires finales seulement si leurs similarités moyennes appartiennent à l'intervalle de validité de la classe adéquate.
— Si $Sim\_moy(Cellule) \in VI$; ajouter $Cellule$ aux cellules mémoires finales. (Cellule= cellule moyenne ou clone muté).

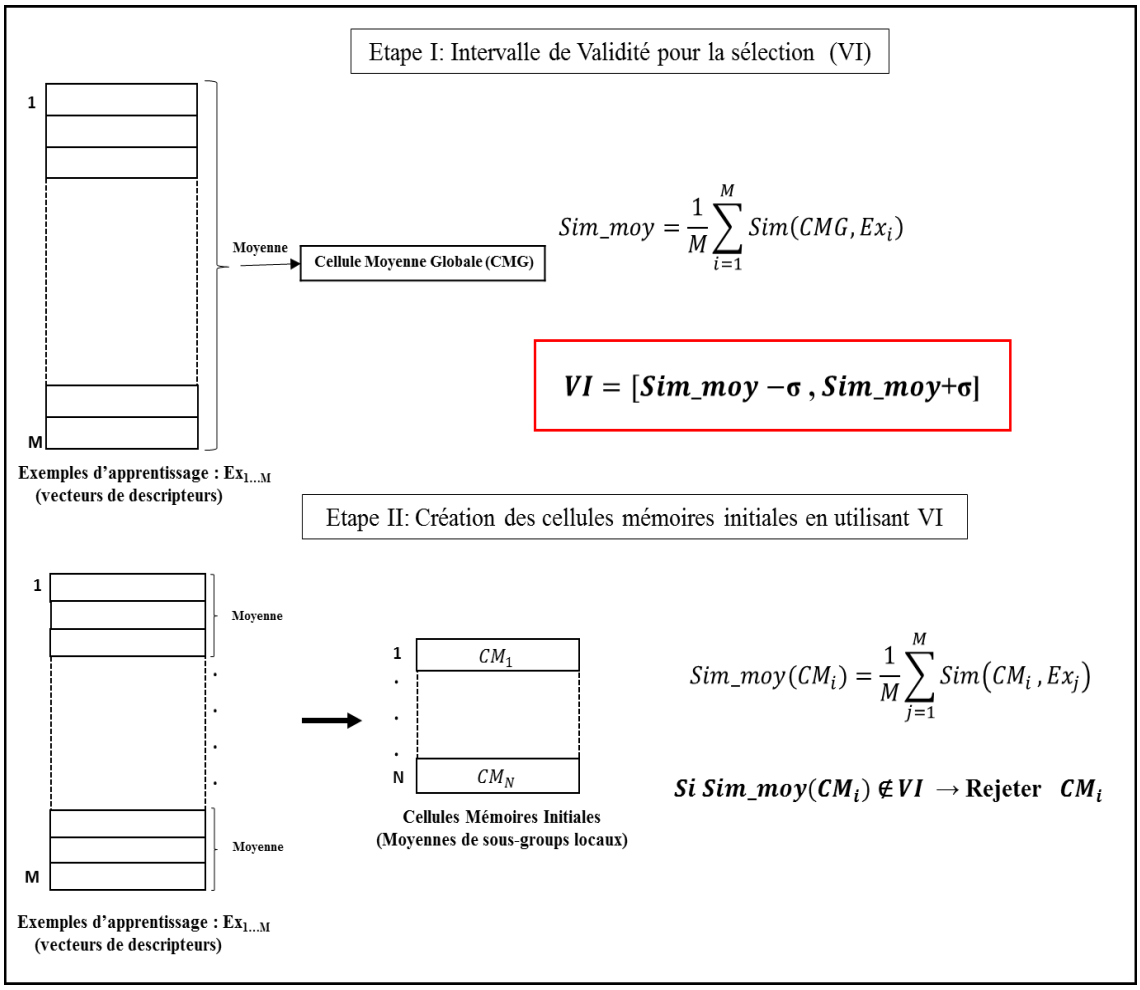


FIGURE 3.8 – Schema de création de l'intervalle de validité (VI) et sélection des cellules mémoires initiales (étapes 1 et 2) ou chaque cellule est représentée par un vecteur de descripteurs.

— Sinon, rejeter Cellule.

En appliquant cette condition supplémentaire dans la sélection, nous permettons la génération d'une population de cellules mémoires généralisées de représentativité globale pour chaque classe. La diversité des données est maîtrisée car chaque cellule mémoire finale présente une similarité moyenne spécifique à sa classe, et aucune cellule d'apprentissage n'est supprimée. En outre, les cellules finales ont la capacité de reconnaître plus facilement les exemples de leurs classes correspondantes. La Figure 3.9 présente le diagramme général de l'algorithme VI-CS.

Nous présentons dans la section 3.4 les résultats d'application des trois approches MG-CLONALG, AC-CLONALG et VI-CS dans la classification du cancer du sein.

3.4 Résultats Expérimentaux

Cette section regroupera l'ensemble des résultats obtenus par l'application des approches MF-CLONALG, AC-CLONAG et VI-CS sur les deux bases de données WDBC et DDSM.

On commencera d'abord par donner les différents paramètres utilisés dans les évaluations, et on discutera les résultats obtenus de chaque méthode par la suite.

3.4.1 Paramètres utilisés

— Étant donné que les bases de données sont normalisées, la mesure de similarité est calculée à travers la distance euclidienne par l'équation suivante :

$$Sim = 1 - \sqrt{\sum_{i=1}^{L}(x_i - y_i)^2} \tag{3.4}$$

où x et y sont respectivement les vecteurs de descripteurs d'un antigène (exemple d'apprentissage) et une cellule mémoire, et L le nombre total de descripteurs de la base de données.

— Le nombre de cellules mémoires initiales de chaque classe (NB_{CM_i}) qui est aussi le nombre des sous groupes locaux est déterminé aléatoirement par l'équation :

$$NB_{CM_i} = round(rand(\frac{M}{3}, \frac{M}{2})) \tag{3.5}$$

avec M nombre d'antigènes (exemples d'apprentissage) de la classe en question.

— Le nombre de clones de chaque cellule mémoire est calculé proportionnellement à sa valeur de similarité par l'équation suivante :

$$NB_{Clones} = round(\beta * Sim) \tag{3.6}$$

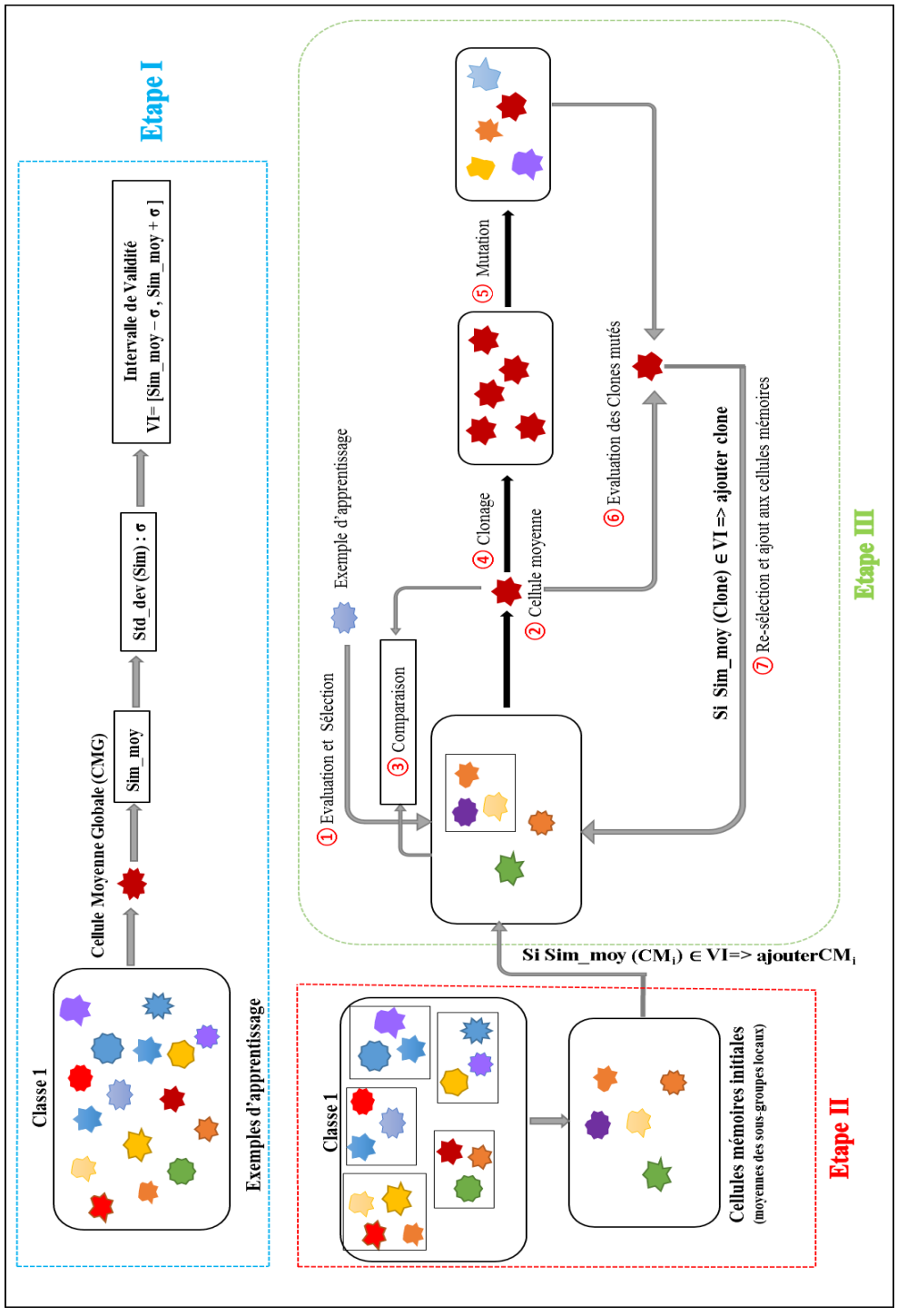


FIGURE 3.9 – Diagramme de VI-CS composé de : Étape I : Création de l'intervalle de validité (VI), Étape II : génération et sélection des cellules mémoires initiales en utilisant (VI) et Étape III : apprentissage du Système Immunitaire Artificiel.

où β est un facteur de clonage fixé par l'utilisateur.

— Afin de fixer la bonne valeur du facteur de clonage β, nous avons procédé à une série de tests en utilisant trois différentes valeurs (3, 5 et 8) pour chacune des deux base de données. Pour 10 générations de l'algorithme, les résultats moyens de 10 exécutions successives sont donnés dans le tableau 3.1. Notons que nous avons utilisé l'approche AC-CLONALG pour fixer le facteur de clonage. Afin de comparer les résultats, nous avons utilisé la même valeur de β pour les trois approches.

β	Résultats de classification (%) sur WDBC	Résultats de classification (%) sur DDSM
3	94.64	94.54
5	**96.80**	**94.98**
8	94.87	93.64

TABLE 3.1 – Résultats de classification de AC-CLONALG sur WDBC et DDSM en utilisant différentes valeurs de β

A partir du tableau 3.1, on remarque que pour les deux bases de données les meilleurs résultats de classification sont obtenus avec un facteur de clonage $\beta = 5$, qui est la valeur moyenne. En effet, le nombre de clones influe directement sur la performance du classifieur, un grand nombre de clones pourrait tromper le résultats de classification, et ralentir l'apprentissage. Tandis-que un nombre réduit de clones est insuffisant pour générer des cellules compétentes. Le facteur de clonage est donc fixé à 5 pour les deux bases de données, ce qui implique qu'une cellule mémoire peut générer 5 clones aux maximum, puisque la similarité est comprise entre 0 et 1.

— L'intervalle de mutation de chaque clone est calculé inversement à sa valeur de similarité, c-à-d, plus la similarité est grande, moins l'intervalle de mutation est large. Ainsi, la valeur de mutation est choisie aléatoirement entre $[Sim-1,1-Sim]$.

Le tableau 3.2 récapitule l'ensemble des paramètres utilisés dans l'évaluation des approches proposées.

Après plusieurs générations, les cellules mémoires générées à la fin de l'apprentissage de chaque algorithme sont utilisées dans la phase de test. La moyenne de dix exécutions successives est considérée comme résultat final de classification (en utilisant le principe de 4-fold cross validation à chaque exécution). La simulation et l'implémentation sont faites en utilisant MATLAB 7.11.0.

Paramètre	Valeur
Similarité (Sim)	1 - Distance Euclidienne
NB_{Clones}	$\beta * Sim$
β (facteur de clonage)	WDBC= 5 ; DDSM=5
Mutation	Rand($Sim - 1$,$1 - Sim$)
NB_{CM_i}	$rand(\frac{M}{3}, \frac{M}{2})$
k (Taille sous-groupe local)	$\frac{M}{NB_{CM_i}}$

TABLE 3.2 – Paramètres utilisés dans l'évaluation des approches MF-CLONALG, AC-CLONALG et VI-CS sur les deux bases de données WDBC et DDSM

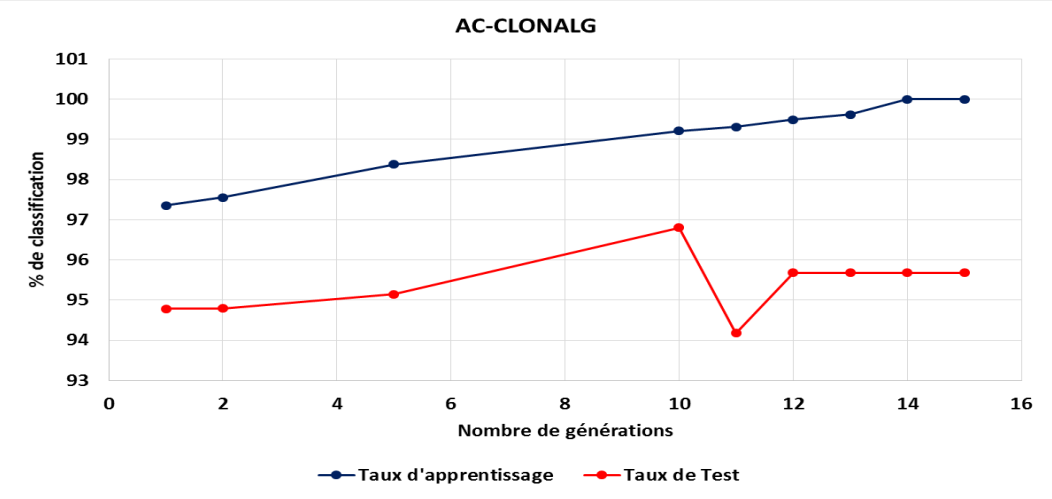


FIGURE 3.10 – Courbes des taux d'apprentissage et de test de l'algorithme AC-CLONALG sur la base WDBC.

Contrairement à la plupart des algorithmes de sélection clonale, les solutions que nous avons proposé dans nos trois approches consistent à créer et ajouter des cellules mémoires sans supprimer ou remplacer aucune cellule. Cela peut avoir un effet sur le temps d'apprentissage, et augmente le risque du sur-apprentissage. On peut vraisemblablement parler de sur-apprentissage (overfitting en anglais) si l'erreur d'apprentissage diminue alors que l'erreur du test augmente de manière significative. Cela signifie que notre programme continue à améliorer ses performances sur les échantillons d'apprentissage mais perd son pouvoir de prédiction sur ceux du test. Pour avoir un programme qui généralise bien, on arrête l'apprentissage dès que l'on observe cette divergence entre les deux courbes.

Dans le but de fixer un nombre maximum de générations permettant d'avoir de bons résultats tout en évitant le sur-apprentissage, nous avons exécuté l'algorithme AC-CLONALG sur la base WDBC pour plusieurs générations allant de 1 à 15. La figure 3.10 trace les courbes des taux d'apprentissage et de test obtenus.

D'après la figure 3.10, on remarque qu'au delà de 10 générations le taux d'apprentissage continue de progresser jusqu'au max (100%), contrairement au taux de test qui commence à diminuer. Par conséquent, nous avons décidé d'arrêter nos exécutions à 10 générations. De cette manière, on évitera le sur-apprentissage ainsi que le ralentissement inutile de

l'algorithme.

3.4.2 Résultats expérimentaux de l'algorithme MF-CLONALG

Dans cette section on détaillera les résultats obtenus par l'application de l'approche MF-CLONAG sur les deux bases de données WDBC et DDSM. Rappelons que la méthode MF-CLONALG (présentée dans la section 3.3.2) consiste à créer des cellules médianes à partir des meilleurs cellules mémoires en vue du clonage et de mutation. L'ensemble de ces paramètres utilisés est regroupé dans le tableau 3.2. Dans ce qui suit, nous allons d'abord présenter les résultats obtenus sur la base WDBC, et après ceux obtenus sur la base DDSM.

3.4.2.1 Résultats sur la Base WDBC

La base de données WDBC est constituée de 569 exemples dont chacun comporte 30 descripteurs. Pour cette raison, la cellule médiane sélectionnée est la 15ème de la matrice des meilleurs cellules mémoires. Les taux d'apprentissage et de test pour différents nombres de générations de MF-CLONALG sur WDBC sont résumés dans le tableau 3.3.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	96.85± 0.58	94.48± 1.34
2	97.35± 0.62	94.86± 1.28
5	98.13± 0.18	94.93± 0.94
10	99.02± 0.19	95.03± 0.50

Table 3.3 – Résultats d'application de MF-CLONALG sur la base WDBC

A partir du tableau 3.3, on remarque que plus on augmente le nombre de générations, plus les résultats d'apprentissage et de test sont meilleurs. De plus, pour 10 générations, le taux d'apprentissage arrive à une moyenne de 99.02% et le taux de test à 95.03%. Par ailleurs, on constate aussi que les écart-types des taux d'apprentissage et de test par rapport à leurs moyennes diminuent avec l'augmentation du nombre de générations. Ce qui implique la bonne précision de l'apprentissage et du test.

3.4.2.2 Résultats sur la Base DDSM

Comme la base de données DDSM est comporte 424 masses, chacune constituée de 22 descripteurs, la cellule médiane sélectionnée est la 11ème de la matrice des meilleurs cellules mémoires.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	93.80± 0.81	93.75± 1.10
2	94.25± 0.72	93.87± 0.86
5	94.53± 0.61	93.98± 0.73
10	95.58± 0.36	94.91± 0.61

TABLE 3.4 – Résultats d'application de MF-CLONALG sur la base DDSM

Les résultats d'application de l'algorithme MF-CLONALG sur la base de données DDSM (tableau 3.4) montrent que l'apprentissage et le test sont meilleurs et plus précis quand le nombre de générations croit. Ce qui est en cohérence avec les résultats du même algorithme sur la base WDBC (tableau 3.3). Néanmoins, on constate que les résultats sur la base DDSM sont moins bons par rapport à ceux de la base WDBC (moyennes d'apprentissage et de test relativement plus petites). Cela peut être expliqué par le nombre d'exemples d'apprentissage de la base DDSM qui est moins important comparé à l'autre base.

3.4.3 Résultats de l'approche AC-CLONALG

Nous présenterons dans cette section les résultats d'application de l'approche AC-CLONALG (section 3.3.3) sur les base de données WDBC et DDSM.

3.4.3.1 Résultats sur la Base WDBC

Dans l'approche AC-CLONALG, l'idée est de sélectionner P cellules mémoires proches de l'antigène (exemple d'apprentissage) pour créer une cellule moyenne potentielle pour le clonage et la mutation. Le problème rencontré à cette étape est dans le choix du nombre de ces cellules. En effet, les questions posées étaient : Combien faut-il sélectionner de cellules mémoires pour créer une cellule moyenne pertinente ? et ce choix influe-t-il sur les résultats ? Pour répondre à ces questions, nous avons effectué une série de tests en utilisant différentes valeurs de P (3,5,10,20 et 50), et comparé les résultats obtenus pour 10 générations. Les taux de test des différentes expérimentations sont donnés dans le tableau 3.5.

P Cellules mémoires	Taux de Test(%)
3	93.42
5	94.39
10	**96.80**
20	95.72
50	94.22

TABLE 3.5 – Résultats d'application de AC-CLONALG sur la base WDBC en utilisant différentes valeurs de P.

D'après le tableau 3.5, il est facilement remarquable que le nombre P de cellules sélectionnées pour générer la cellule moyenne a une influence sur les résultats de l'algorithme. En effet, les résultats de AC-CLONALG avec P=10 cellules sont meilleurs que ceux avec P=3 ou P=20 ou P=50. Cela veut dire qu'il faut assez de cellules pour générer une cellule mémoire moyenne pertinente, mais pas énormément de cellules au point de générer des cellules moins bonnes que la cellule la plus proche de l'antigène.

Le tableau 3.6 présente les résultats de classification de AC-CLONALG sur la base la base WDBC pour différents nombres de générations avec P=10.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	97.35± 0.73	94.78± 1.29
2	97.55± 0.34	94.79± 1.00
5	98.38± 0.41	95.15± 0.91
10	99.21± 0.29	96.80± 0.52

TABLE 3.6 – Résultats d'application de AC-CLONALG sur la base WDBC

Après 10 générations de l'algorithme, on arrive à un taux d'apprentissage de 99.21 % ; et un taux de test de 96.80% avec des déviations de 0.29 et 0.52 respectivement. On peut constater que cette approche améliore les résultats de classification mieux que MF-CLONALG sur WDBC.

3.4.3.2 Résultats sur la Base DDSM

Comme sur la base de données précédente, afin de choisir le nombre de cellules mémoires à sélectionner pour créer la cellule à moyenne, nous avons appliqué l'algorithme AC-CLONALG en utilisant différentes valeurs de P. Puisque la base DDSM est moins consistante que la précédente, il est impossible de sélectionner jusqu'à 50 cellules. Nous nous sommes arrêté donc à 20 cellules mémoires. Les résultats de chaque expérimentation (pour 10 générations) sont donnés dans le tableaux 3.7.

P Cellules Mémoires	Taux de Test(%)
3	94.40
5	**94.98**
10	94.11
20	93.20

TABLE 3.7 – Résultats d'application de AC-CLONALG sur la base DDSM en utilisant différentes valeurs de P

D'après le tableau 3.7, le meilleur résultat est obtenu avec une valeur de P=5 , qui est une valeur moyenne par rapport à la consistance de la base de données. les résultats de classification de la base DDSM en utilisant cette valeur sont présentés dans le tableau 3.8.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	95.00± 0.62	93.76± 1.39
2	95.36± 0.85	93.87± 1.18
5	95.79± 0.46	94.09± 0.93
10	96.52± 0.25	94.98± 0.78

TABLE 3.8 – Résultats d'application de AC-CLONALG sur la base DDSM

Sur la base DDSM ; le meilleur résultat est obtenu après 10 générations. Le taux d'apprentissage est de 96.52% avec une déviation de 0.25, et le taux de test est 94.98% avec une déviation de 0.78.

Comme pour l'approche précédente (MF-CLONALG), les résultats d'apprentissage et de test de AC-CLONALG sur les deux bases de données sont meilleurs et plus précis avec l'augmentation du nombre de générations de l'algorithme. On remarque aussi que les résultats obtenus sur la base WDBC sont meilleurs que ceux obtenus sur la base DDSM, pour les raisons évoquées auparavant (base WDBC plus consistante que la base DDSM en termes de nombre d'exemples et nombre de descripteurs).

Cependant, les résultats de l'approche AC-CLONALG sur les deux bases de données se montrent légèrement meilleurs que les résultats de l'approche MF-CLONALG. on peut donc dire que l'idée de création d'une cellule moyenne est meilleure que celle de créer une cellule médiane. En effet, la création d'une cellule moyenne implique forcément l'utilisation de la totalité des P cellules sélectionnées pour sa création. Tandis-que la génération de la cellule médiane est basée sur le tri des descripteurs des meilleures cellules mémoires. Il est alors possible de générer une cellule médiane à partir d'une ou deux cellules, car elle(s) peut(vent) contenir l'intégralité des valeurs médianes des attributs de la base de données.

3.4.4 Résultats de l'approche VI-CS

Après avoir présenté les résultats des deux premières approches proposées pour améliorer la sélection clonale, nous présentons dans cette section les résultats d'application de la méthode VI-CS (Validity Interval Clonal Selection) proposée dans la section 3.3.4. L'algorithme VI-CS vise à renforcer la diversité de l'algorithme AC-CLONALG présenté dans la section 3.3.3. En effet, pour améliorer la représentativité globale des données d'apprentissage et préserver une bonne diversité dans l'algorithme, nous avons proposé d'utiliser un intervalle de validité pour chaque classe à apprendre en nous basant sur les caractéristiques statistiques de ces dernières.

Les résultats obtenus par la méthode VI-CS sur les bases de données WDBC et DDSM sont respectivement présentés dans les tableaux 3.9 et 3.10.

3.4.4.1 Résultats sur les Bses WDBC et DDSM

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	96.96 ±0.85	94.18 ±1.32
2	97.80 ±0.30	94.75 ±1.05
5	98.55 ±0.32	95.48 ±0.90
10	98.95 ±0.16	97.58 ±0.22

TABLE 3.9 – Résultats d'apprentissage et de test de l'algorithme VI-CS sur la base de données WDBC

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)
1	94.55 ±1.40	94.64 ±0.87
2	94.70 ±1.60	94.91 ±1.12
5	95.54 ±0.56	95.18 ±0.80
10	96.66 ±0.20	95.76 ±0.39

TABLE 3.10 – Résultats d'apprentissage et de test de l'algorithme VI-CS sur la base de données DDSM

Les résultats d'application de VI-CS sur les deux bases WDBC et DDSM sont les mêmes que celles des approches précédentes concernant les points suivants :

— Les taux d'apprentissage et de test croient avec l'augmentation du nombre de générations de l'algorithme.
— L'augmentation du nombre de générations contribue aussi à améliorer la précision des résultats de l'algorithme en réduisant la valeur de l'écart-type.
— Les résultats obtenus sur la base WDBC sont légèrement meilleurs que ceux obtenus sur la base DDSM, car cette dernière est moins consistante.

Puisque VI-CS est une amélioration de l'algorithme AC-CLONALG, nous présentons dans la figure 3.11 une comparaison entre les deux approches en termes de similarités moyennes des deux classes Bénigne et Maligne de la base WDBC (car elle est plus consistante que la base DDSM). Les similarités moyennes des cellules mémoires finales obtenues par les deux approches sont calculées et comparées à la similarité moyenne des cellules originales de la base.

D'après la figure 3.11, on remarque que la similarité moyenne des cellules mémoires obtenues par l'approche VI-CS (0.11 pour la classe maligne et 0.10 pour la classe bénigne) est plus proche de celle des cellules originales de la base (0.14 pour la classe maligne et 0.13 pour la classe bénigne). On peut donc dire que l'algorithme VI-CS a contribué à préserver la diversité de l'algorithme AC-CLONALG. En effet, l'utilisation de l'intervalle de validité

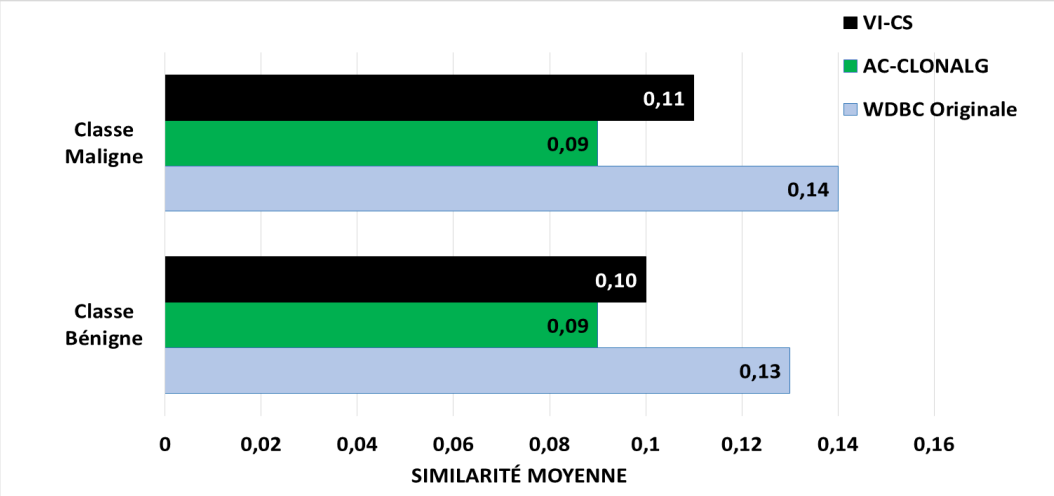


FIGURE 3.11 – Comparaison entre les valeurs de similarités moyennes de la base WDBC et les cellules mémoires finales obtenues par les algorithmes AC-CLONALG et VI-CS.

pour chaque classe a permit de sélectionner les cellules mémoires les plus représentatives des données d'apprentissage. Cela explique pourquoi des résultats de classification de VI-CS sont meilleurs que ceux de AC-CLONALG.

Une comparaison des résultats des trois approches proposées avec les travaux de la littérature sera donnée dans la section 3.4.5.

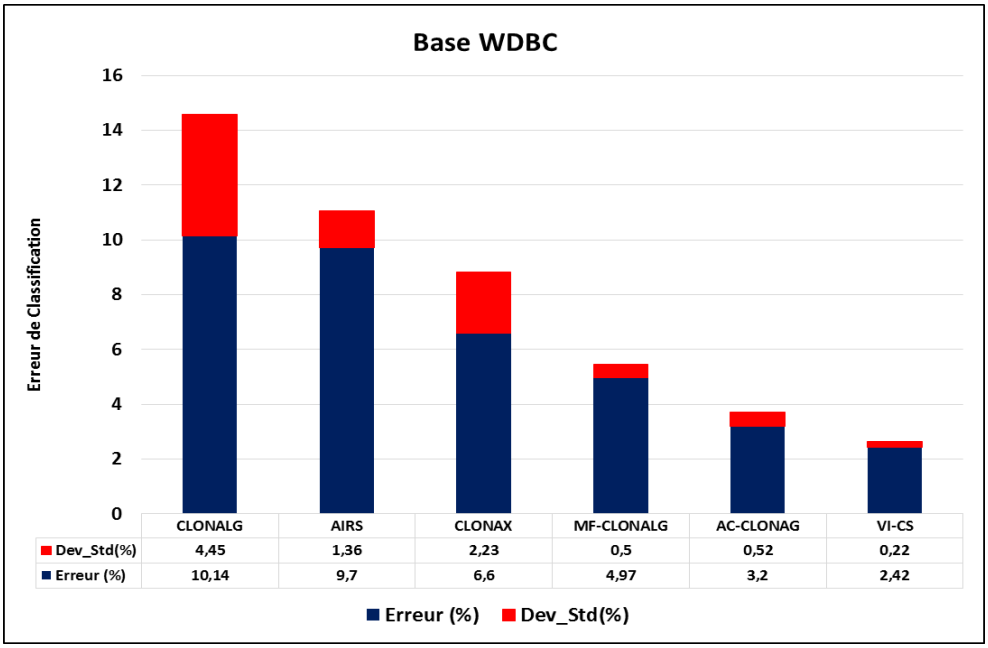
3.4.5 Étude comparative

Dans cette partie du chapitre, on établie une étude comparative entre les trois approches que nous avons proposé, et quelques algorithmes de sélection clonale de la littérature que nous avons implémenté. Nous avons appliqué chaque algorithme sur les deux bases de données (WDBC et DDSM) en utilisant les mêmes paramètres cités dans le tableau 3.2. Les résultats de chaque application pour 10 générations sont cités dans le tableau 3.11. Les histogrammes des erreurs de classification de chaque approche sont illustrés dans la figure 3.12.

Le tableau 3.11 illustre l'efficacité des améliorations de l'algorithme CLONALG. On remarque que les trois approches proposées présentent toujours les meilleurs résultats de classification sur les deux bases de données avec un écart-type plus faible.

Les principales contributions que nous avons apporté et qui permettent de justifier l'amélioration des résultats obtenus par les algorithmes : CLONALG [DCVZ02], AIRS [Wat01] et CLONAX [SS11] sont :

— Amélioration de l'initialisation par la création de cellules mémoires initiales spécifiques à chaque classe d'apprentissage.
— Aucune cellules (mémoire ou originale) n'est supprimée ou remplacée.



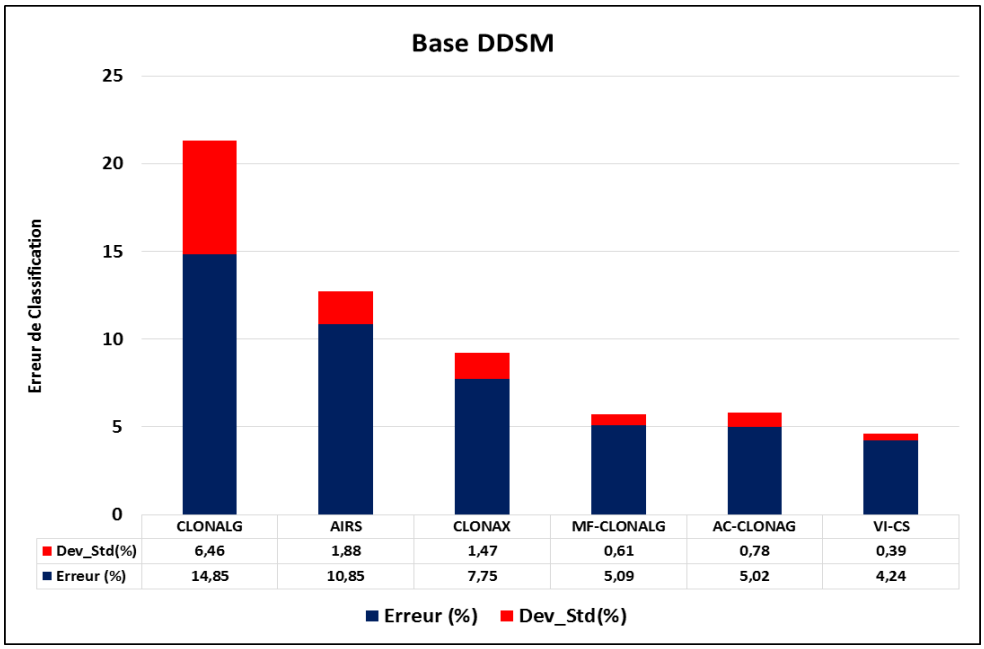


FIGURE 3.12 – Histogrammes des erreurs de classification sur la base WDBC(haut) et la base DDSM (bas)

Algorithme	(%) WDBC	(%) DDSM
CLONALG	89.86± 4.45	85.15± 6.46
AIRS	90.3± 1.36	89.15± 1.88
CLONAX	93.4± 2.23	92.25± 1.47
MF-CLONALG	**95.03± 0.50**	**94.91± 0.61**
AC-CLONALG	**96.80± 0.52**	**94.98± 0.78**
VI-CS	**97.58± 0.22**	**95.76± 0.39**

TABLE 3.11 – Comparaison des résultats des trois approches proposées avec des méthodes de classication connues : AIRS, CLONALG et CLONAX.

— Aucune cellule aléatoire n'est introduite à l'algorithme.
— Création de cellules mémoires potentielles médianes (MF-CLONALG) ou moyennes (AC-CLONALG , VI-CS) pour le clonage et la mutation.
— Utilisation d'un intervalle de validité de sélection (VI-CS) afin d'améliorer la représentativité globale et préserver une bonne diversité des données d'apprentissage.

Même si les approches proposées ont nettement amélioré les résultats de classification des algorithmes de sélection clonale, nous avons pu constater qu'elles nécessitent des coûts de calcul non négligeables. En effet, le fait d'ajouter des cellules mémoires à chaque présentation d'un antigène pour plusieurs générations permet d'améliorer la reconnaissance, mais en contre partie cela nécessite un temps d'apprentissage plus important.

3.5 Conclusion

Nous nous sommes intéressé dans ce chapitre à l'amélioration d'un algorithme de base dans la famille des algorithmes de sélection clonale artificielle. Tous d'abord, nous avons commencé par la présentation de l'algorithme CLONALG, puis nous avons cité quelques remarques que l'on a constaté sur ce dernier. Ces remarques concernent la façon dont il est initialisé, et l'introduction d'individus aléatoires pour maintenir la diversité. Par la suite, nous avons présenté diverses solutions pour traiter ces remarques. Pour améliorer l'étape d'initialisation dans CLONALG, nous avons proposé de créer des cellules mémoires initiales spécifiques à chaque classe d'apprentissage à partir de sous-groupes locaux de ces dernières.

Concernant le problème d'introduction de cellules aléatoires à l'algorithme, nous avons présenté trois différentes méthodes nommées MF-CLONALG (Median Filter Clonal Algorithm) , AC-CLONALG(Average Cells Clonal Algorithm) et VI-CS (Validity Interval Clonal Selection algorithm). Chacune de ces approches a pour but de créer des cellules mémoires potentielles et préserver la diversité dans l'algorithme sans avoir a rejeter aucune cellule qui pourrait être pertinente dans les générations suivantes. De plus, la méthode VI-CS introduit la notion d'intervalle de validité pour une représentativité convenable des classes d'apprentissage.

Les résultats obtenus sur les deux bases de données WDBC et DDSM sont meilleurs comparés à ceux de CLONALG, l'algorithme sur lequel les améliorations ont été faites,

aussi que ceux d'autres algorithmes classiques de sélection clonale.

L'efficacité des algorithmes proposées sera meilleure si le temps d'apprentissage pourra être optimisé. Dans le chapitre suivant, on présentera une méthode d'accélération de l'apprentissage de ces algorithmes tout en améliorant leur précision de classification.

Chapitre 4

OPTIMISATION DU TEMPS D'APPRENTISSAGE PAR CATÉGORISATION LOCALE

4.1 Introduction

Le but de ce chapitre est de trouver une solution pour optimiser le temps nécessaire à l'apprentissage des approches proposées dans le chapitre précédent, sans réduction de leur performance. Pour cela, nous avons proposé une méthode basée sur trois algorithmes : K-Means, le Réseau à Fonctions de base Radiales (RBF) et le Système Immunitaire Artificiel (SIA). L'approche proposée vise à réduire le temps de calcul des algorithmes de classification SIA sans affecter leur performance. Le principe de cette approche consiste à partitionner les ensembles de cellules mémoires en plusieurs catégories (clusters) locales en utilisant l'algorithme K-Means, et l'apprentissage de chaque catégorie par le réseau de neurones RBF. Le but de la catégorisation locale des données est de réduire le nombre de tests à effectuer par chaque exemple d'apprentissage dans les algorithmes SIA pour sélectionner la cellule la plus proche à cloner qui améliore la reconnaissance des cellules.

Les résultats obtenus sur les bases de données WDBC et DDSM montrent l'efficacité du classifieur proposé, que ce soit dans la précision de la classification ou les coûts de calcul, par rapport aux algorithmes proposés dans le chapitre 3.

4.2 Complexité des algorithmes de sélection clonale

La complexité d'un programme informatique se réfère à la quantité de temps requis pour l'exécuter dans le pire des cas [HS78]. On peut mesurer le temps requis par un algorithme en comptant le nombre maximal d'instructions exécutées, qui est proportionnel au nombre maximum de fois que chaque boucle est exécutée. Concernant les algorithmes de sélection clonale, nous avons vu dans le chapitre précédent qu'ils comportent trois principales étapes de traitement qui sont :

— Le calcul des similarités (des cellules mémoires ou des clones mutés) ;
— La sélection (et la re-sélection) des meilleures cellules mémoires ou clones ;

— Le clonage et la mutation des clones ;

Ce qui veut dire que le temps requis pour une génération d'un algorithme de sélection clonale dépend essentiellement de ces trois étapes.

Prenons l'exemple de CLONALG, les auteurs ont étudié sa complexité dans [DCVZ02]. Ils ont utilisé des paramètres qui caractérisent les calculs effectués, tels que la taille de la population des cellules mémoires N, le nombre de clones NB_{Clones}, la taille de la base d'apprentissage M, et la dimension des données L (nombre de descripteurs de la base de données).

La sélection des meilleures cellules mémoires pour le clonage se fait par le tri du vecteur des similarités (F), et l'extraction des n éléments correspondants aux premiers éléments du vecteur F, ce qui peut être effectué en un temps $O(N)$. Aussi, le temps de calcul des étapes de sélection et de re-sélection est de l'ordre $O(N)$ et $O(NB_{Clones})$. L'étape de mutation des clones demande un temps de calcul de l'ordre $O(NB_{clones} * L)$. Sachant que ces étapes doivent être effectuées pour chaque exemple de la base d'apprentissage, le temps d'exécution de l'algorithme CLONALG peut être calculé en sommant le temps de calcul de chacune de ces 3 étapes, et en multipliant le résultat par M (nombre d'exemples d'apprentissage de toutes les classes). Le temps requis pour une génération de CLONALG est donc de l'ordre : $O(M * (N + NB_{Clones} * L)$.

A.Sharma et D.Sharma ont aussi étudié la complexité de leur algorithme de sélection clonale pour la classification. Le temps de calcul de CLONAX est de l'ordre $O(M^2 + M * NB_{Clones} * L)$ qui est le même que celui de CLONALG pour la reconnaissance des formes [SS11] (au pire des cas N peut être égal à M).

Il ne faut pas oublier de mentionner que ces calculs sont pour une seule génération, et qu'il en faut plusieurs pour converger vers une population de cellules mémoires pertinentes. Il est important aussi de préciser que la taille de l'ensemble des cellules mémoires de chaque classe augmente à chaque présentation d'un antigène par le nombre de clones ajoutés (cela dépend de l'algorithme) ce qui nous amène à une population de cellules mémoires d'une taille importante qui nécessite un temps conséquent de calcul.

Plusieurs articles abordent le sujet de l'apprentissage lent des algorithmes de sélection clonale. Dans [WG03] les auteurs concluent que la faiblesse de CLONALG est le temps nécessaire pour générer une bonne population de cellules mémoires. Ils ont proposé CLONACLAS, un algorithme avec la même complexité de CLONALG mais qui converge plus rapidement. Dans leur évaluation des Systèmes Immunitaires Artificiels, les auteurs de [Gar05] parlent de la lenteur de CLONALG et AIRS à cause des différentes évaluations effectuées sur les cellules mémoires et les clones, pour la sélection et la re-sélection des meilleurs individus. Le travail proposé dans [WT04a] améliore l'algorithme AIRS, la suppression de quelques étapes inutiles de l'algorithme pour améliorer sa précision à conduit à une réduction du temps de calcul de l'algorithme. Une autre version de l'algorithme AIRS conçue pour la distribution à travers un nombre variable de processus a été proposée dans [WT04b]. L'approche de parallélisme dans AIRS était simple, impliquant des étapes ajoutées au schéma d'apprentissage standard de l'algorithme. Ces étapes consistent à répartir l'ensemble d'apprentissage de N partitions et l'allocation de chaque partition à

un processus d'apprentissage. Après la génération de N populations de cellules mémoires à partir de chaque partition, un schéma de fusion est utilisé pour créer une population générale de cellules mémoires pour la classification. Les résultats ont montré que tant que l'ensemble de données d'apprentissage n'est pas partitionné trop largement, une accélération peut être observée en exécutant AIRS en parallèle. Les auteurs ont aussi remarqué qu'il y a eu une certaine perte dans les avantages de AIRS et que les résultats étaient peu concluants.

Nous avons présenté dans le chapitre précédent des approches améliorant l'algorithme de classification CLONALG [DCVZ02] en particulier, et les algorithmes de sélection clonale en général. Ces méthodes consistent à créer des cellules efficaces de représentativité globale des classes à apprendre. A la fin d'apprentissage de chaque exemple (antigène), ces approches ajoutent les meilleurs clones mutés à l'ensemble des cellules mémoires finales de leurs classes correspondantes. Ce processus a contribué efficacement à améliorer la précision des algorithmes de sélection clonale artificielle. Cependant, il nécessitent un temps important pour pouvoir traiter convenablement toutes les cellules mémoires dont la taille augmente à chaque génération.

Dans ce chapitre, nous présenterons une méthode d'optimisation du temps d'apprentissage des algorithmes de sélection clonale, tout en préservant leur précision. En effet, contrairement aux travaux d'accélération de l'apprentissage qui ont causé une perte dans la performance du classifieur tels que [WT04b], l'approche que l'on propose a deux objectifs : le premier est la rapidité de l'apprentissage, et le second est celui de préserver ou améliorer la performance de l'algorithme.

4.3 Catégorisation Locale de Base de données (LDC-AIS)

Cette section présentera les différents principes et détails de l'approche que nous avons proposé, ainsi que les méthodes et algorithmes utilisé. On commencera par donner le principe de la méthode, les diverses étapes de l'algorithme sont détaillées par la suite.

4.3.1 Principe de l'approche

L'endroit logique pour chercher une façon de résoudre le problème de l'apprentissage lent des algorithmes de sélection clonale est le Système Immunitaire Naturel (SIN) lui même. En effet, le SIN est par nature rapide, l'observation de ce dernier pourrait fournir un aperçu sur la façon d'aborder le problème. On a vu dans le chapitre 2, que la réponse immunitaire commence par un contact avec l'antigène et l'activation des cellules spécifiques à ce dernier pour le neutraliser.

L'idée de l'approche que l'on propose vient de ce principe. L'activation des cellules mémoires spécifiques à l'antigène seulement pour participer à la réponse immunitaire a attiré notre attention. Effectivement, l'activation de tout l'ensemble des cellules mémoires de la classe pour sélectionner une cellule à cloner consomme énormément de temps, d'autant

plus que cet ensemble s'élargit avec chaque passage d'un antigène. Il faut donc réduire la taille des cellules mémoires activées pour répondre à un antigène en prenant seulement la partie spécifique à ce dernier.

Nous avons donc décidé de catégoriser (partitionner) localement les cellules mémoires initiales de chaque classe d'apprentissage, et apprendre chaque catégorie à part. L'apprentissage d'un individu (antigène) se fera en le comparant seulement avec sa catégorie spécifique de cellules mémoires. De cette manière, on réduira les calculs nécessaires pour apprendre chaque exemple sans aucun effet sur les résultats de classification.

4.3.1.1 Méthode et algorithmes utilisés

Pour partitionner les ensembles de cellules mémoires de chaque classe en plusieurs catégories locales, nous avons choisi d'utiliser la méthode de classification non-supervisée la plus communément utilisée et la plus simple qui est celle de l'algorithme K-Means.

La méthode discriminative de K-Means [Mac67] utilise un principe simple et naturel qui consiste à diviser M points de dimension N en k groupes (clusters) $C_{1..k}$. Chaque groupe dispose d'un centre unique, de telle manière que chaque point est associé au groupe dont le centre est le plus proche. Pour calculer le centre le plus proche de chaque point, K-means utilise des métriques de distance, et comme dans notre travail on utilise la distance euclidienne comme mesure de similarité, nous l'avons choisi de l'utiliser comme métrique de distance dans K-means.

De point de vue algorithmique, l'algorithme K-means effectue un processus itératif qui alterne :

1. L'affectation de chaque objet $O \in M$ au cluster C_i de centre P_i tel que $Dist(O, P)$ est minimale.
2. Recalculer P_i de chaque cluster (catégorie).

Notons que le nombre k de catégories ainsi que les centres de ces dernières sont fixés préalablement. La figure 4.1 retrace le principe de l'algorithme K-means. pour une description plus détaillée de l'algorithme se référer à [HW79].

Après avoir partagé les classes d'apprentissage en plusieurs catégories par l'algorithme K-means, on procède à un apprentissage de chaque catégorie pour faciliter le choix des cellules mémoires spécifiques à l'antigène par la suite.

Comme le but est d'accélérer l'apprentissage des algorithmes de sélection clonale, il faut choisir un algorithme d'apprentissage rapide et efficace, avec un minimum de paramètres à régler. Notre choix a vite été orienté vers les Réseaux de Neurones Artificiels (RNA) ou les Séparateurs à Vaste Marge (SVM).

Parmi les différents modèles connexionnistes existants, les deux réseaux les plus utilisés et les plus pertinents sont le Perceptron Multi-Couches (PMC) en Anglais Mlti-Layer Perceptron (MLP), et les Réseaux à Fonctions de base Radiales en anglais Radial Basis Function (RBF). Notre choix entre ces deux réseaux de neurones a été vite fixé, car le MLP à un inconvénient majeur qui est sa lente convergence, d'autant plus qu'il nécessite

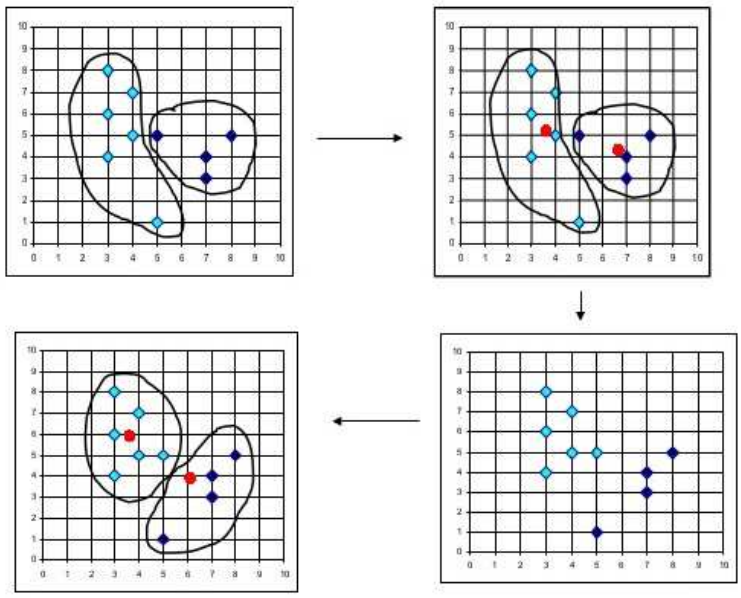


FIGURE 4.1 – Principe de K-Means

plus de paramètres à régler que le RBF qui est constitué d'une seule couche cachée dont il faut choisir le nombre de neurones.

Pour choisir entre le réseau RBF et les SVM, nous avons procédé à un test sur nos deux bases de données et le SVM a été moins performant. Notre choix pour apprendre les catégories définies par K-means a été donc le réseau de neurones RBF.

Le réseau à fonctions de base radiales [Pow87], connu sous l'abréviation RBF fournit un outil alternatif à l'apprentissage des réseaux de neurones artificiels. L'idée principale est de concevoir un réseau avec une bonne capacité de généralisation et d'un nombre minimum de nœuds pour éviter des longs calculs inutiles par opposition au MLP [MAC+92].

L'architecture du réseau RBF est constituée de trois couches : la couche d'entrée, une seule couche cachée qui contient les neurones, et la couche de sortie. Chaque couche est complètement liée à la suivante. La couche cachée est composée d'un certain nombre de neurones avec des fonctions d'activation radiales qui sont généralement des gaussiennes. L'architecture du réseau RBF est montrée dans la figure 4.2.

Une fonction de base radiale est une fonction F_i symétrique autour d'un centre C_i $(1 \leq i \leq N)$.

$$F_i(x) = F(\|x - C_i\|) \tag{4.1}$$

où $\|.\|$ désigne une norme.

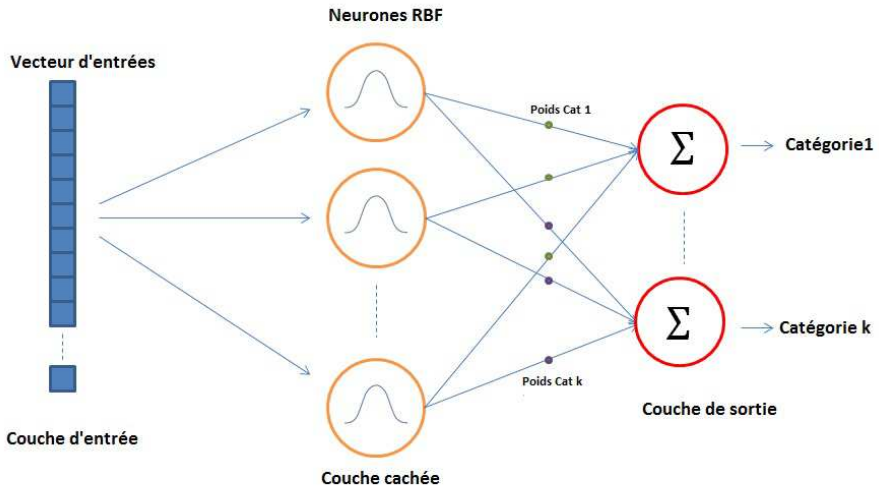


FIGURE 4.2 – Simple architecture du réseau de neurones RBF

La fonction gaussienne (RBF avec la norme euclidienne), est la plus utilisée parmi les fonctions à base radiales. L'équation 4.1 s'écrit alors :

$$F_i(x) = exp(-\frac{(\|x - C_i\|)^2}{\sigma_i^2}) \tag{4.2}$$

où σ_i est la largeur de la i^{ème} unité RBF. La sortie j du réseau RBF est donc :

$$Y_j(x) = \sum_{i=1}^{N}(F_i(x).W_{j,i}) \tag{4.3}$$

avec $W_{j,i}$ le poids de la i^{ème} sortie et $F_0 = 1$.

Les modèles RBF sont particulièrement bien adaptés à l'application des problématiques concrètes dans différents domaines de la recherche scientifique tels que : le traitement de signal, la robotique, la reconnaissance des formes, la classification,etc.

Dans notre application, un modèle RBF est construit pour chaque classe, afin d'apprendre les différentes catégories des cellules mémoires prédéfinies par K-Means. Le but étant de limiter les différents tests à effectuer par les exemples d'apprentissage, et permettre l'accélération des algorithmes de sélection clonale.

4.3.2 Algorithme LDC-AIS

Dans cette partie, on donnera les différents détails et principes de l'approche que l'on propose. Le but principal de ce travail est l'accélération de l'apprentissage des algorithmes de sélection clonale en réduisant le nombre de calculs effectués, le deuxième objectif est celui de conserver ou améliorer leur précision. L'algorithme de Catégorisation locale de

Base de données, Local Database Categorization en anglais (LDC-AIS) est composé de deux phases principales : la phase d'initialisation et la phase d'apprentissage du SIA.

4.3.2.1 Initialisation

La phase d'initialisation dans un algorithme permet de préparer les données pour garantir un bon apprentissage, et dans notre travail, un apprentissage rapide et efficace. Dans les algorithmes de sélection clonale, l'initialisation consiste à construire une population de cellules mémoires initiales pour chaque classe. Cela est fait généralement en sélectionnant des exemples aléatoirement à partir des données d'apprentissage. Dans notre approche, l'initialisation comprends trois étapes :

a) Création des cellules mémoires initiales :

La première étape est la création des cellules mémoires initiales. On a vu dans le chapitre précédent que le choix de ces cellules à une importante influence sur la pertinence de l'apprentissage, car c'est à partir de ces cellules initiales que vont naître les cellules mémoires finales. La création des cellules mémoires initiales pour chaque classe est faite en appliquant la méthode décrite dans le chapitre 2 (Section 3.3.1), c'est à dire par moyennes de sous-groupes locaux.

b) Catégorisation par K-means :

On a expliqué dans la section précédente que le système immunitaire naturel répond rapidement aux antigènes, car seules les cellules spécifiques à ce dernier sont activées pour déclencher une réponse immunitaire. Pour réduire le nombre de calculs effectués par chaque exemple d'apprentissage pour sélectionner la cellule à cloner, la deuxième étape de l'initialisation consiste à partitionner les ensembles de cellules mémoires initiales créées à l'étape précédente .

En effet, en appliquant le même principe que le système immunitaire naturel, on assurera une importante réduction du temps de calcul. Pour cela, les cellules mémoires initiales de chaque classe sont partitionnées en plusieurs catégories selon leurs similarités en utilisant l'algorithme K-Means. Le choix du nombre de catégories de chaque classe est expliqué dans la section 4.4.1.

c) Apprentissage des catégories par RBF :

La dernière étape de l'initialisation est l'apprentissage des catégories. Un modèle RBF est construit pour chaque classe d'apprentissage (dans notre cas un modèle RBF pour la classe bénigne et un autre pour la classe maligne). Les catégories sont apprises par le réseau adéquat comme catégories spécifiques à des antigènes de la même classe. Le rôle du classifieur RBF est de mémoriser chaque catégorie de cellules mémoires pour pouvoir décider pendant l'apprentissage du SIA à laquelle appartient l'antigène, afin d'éviter de le comparer à la totalité des cellules mémoires de la même classe.

Un schéma explicatif des trois étapes d'initialisation est donné dans la figure 4.3.

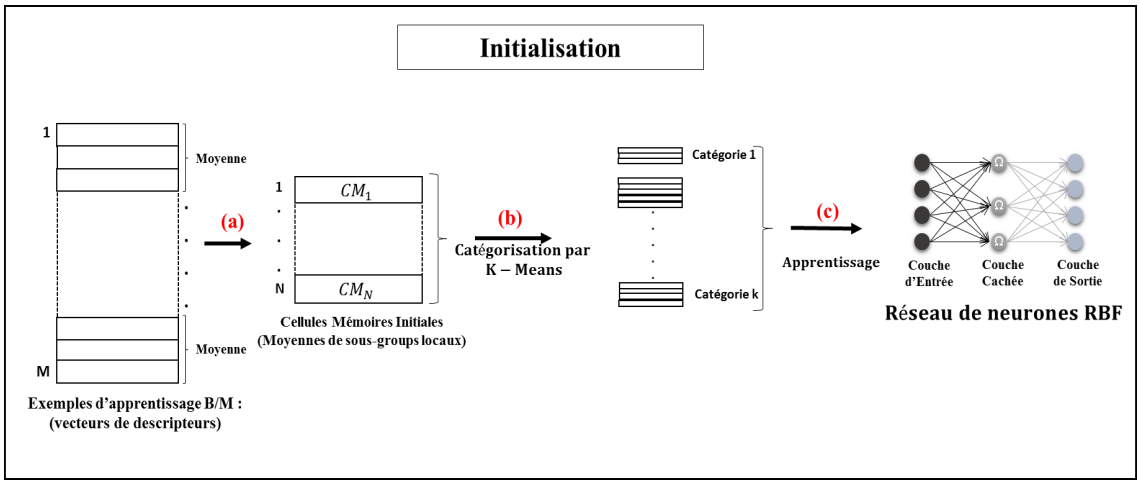


FIGURE 4.3 – Schéma de l'étape d'initialisation de l'algorithme LDC-AIS composé de : a) création des cellules mémoires initiales, b) catégorisation par K-means et c) apprentissage des catégories par le réseau de neurones RBF.

A la fin de l'initialisation, on dispose de populations de cellules mémoires initiales dont chacune est partitionnée en plusieurs catégories apprises par un réseau RBF, pour répondre spécifiquement et plus rapidement aux antigènes.

4.3.2.2 Apprentissage du Système Immunitaire Artificiel

L'apprentissage du SIA commence après l'initialisation des données. Les étapes de l'apprentissage de l'algorithme LDC-AIS sont illustrées dans la figure 4.4. Chacune de ces étapes (de 1 à 4) est exécutée pour un nombre définit de générations (NB_{Gen}) et pour chaque exemple d'apprentissage (de chaque classe) :

1. Test par RBF :

Chaque exemple d'apprentissage est testé par le modèle RBF correspondant à sa classe, le but étant de déterminer à quelle catégorie il appartient.

2. Évaluation de la similarité et sélection :

Après avoir déterminé la catégorie de l'exemple d'apprentissage, l'évaluation de ce dernier est effectuée en calculant sa similarité avec les cellules appartenant à la même catégorie seulement, au lieu de le comparer à tout l'ensemble des cellules mémoires de la même classe. Après l'évaluation des cellules mémoires, nous avons appliqué deux méthodes de sélection :

i) La première consiste à sélectionner une seule cellule mémoire maximisant la valeur de similarité avec l'exemple d'apprentissage pour passer à l'étape 4.3.2.2.

ii) La seconde méthode de sélection suit le principe de l'approche AC-CLONALG proposée au chapitre 3 (section 3.3.3). Au lieu de choisir une seule cellule pour le clonage,

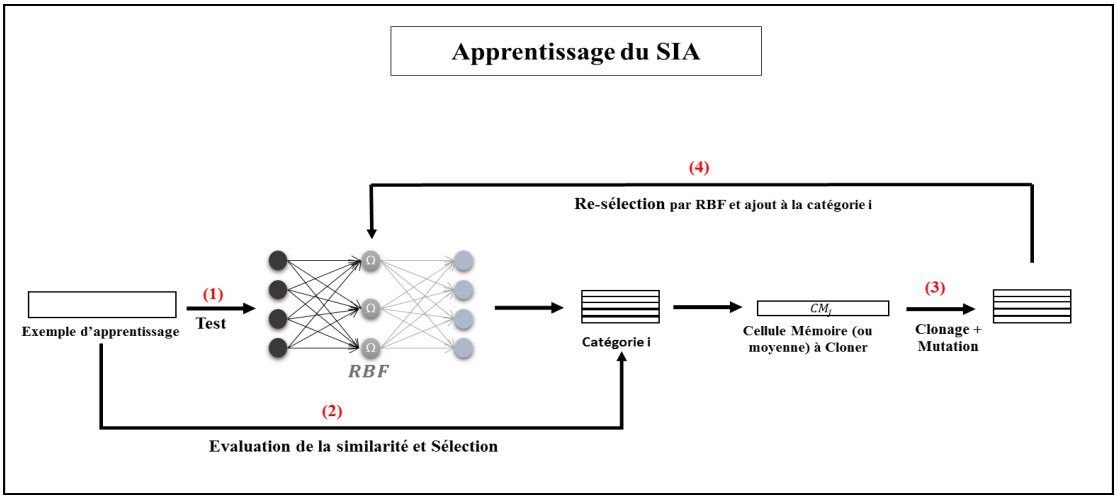


FIGURE 4.4 – Schéma de l'apprentissage de l'algorithme LDC-AIS composé de : 1) test de la catégorie de l'exemple d'apprentissage par RBF, 2) évaluation de la similarité et sélection de la cellule à cloner, 3) clonage et mutation, et 4) Re-sélection des meilleurs clones par le réseau RBF et ajout aux cellules mémoires.

l'exploration de la population des cellules mémoires est effectuée. P cellules mémoires ayant les plus grandes similarités sont sélectionnées, et une cellule moyenne est créée à partir de ces dernières. Comme nous l'avons détaillé au chapitre précédent, la cellule moyenne n'est sélectionnée pour le clonage que si elle présente une meilleure similarité avec l'antigène que la cellule mémoire la plus proche de ce dernier.

Mais dans cette approche, une autre condition est ajoutée. En effet, la cellule moyenne est prise en compte si elle réussit le test par RBF. Si le réseau RBF classe la cellule moyenne dans une autre catégorie que celle de l'exemple d'apprentissage, elle est automatiquement rejetée. Sinon elle est sélectionnée pour passer à l'étape suivante.

Si la classe d'apprentissage comprend X exemples (antigènes) et Y cellules mémoires, cette procédure diminue le nombre de comparaisons de cette étape de $X.Y.[(k-1)/k]$ pour chaque génération. Ce qui nous conduit en tout à une réduction de $G.X.Y.[(k-1)/k]$ comparaisons pour la totalité des générations. Cela implique une réduction importante du temps de calcul pour chaque classe d'apprentissage.

3. Clonage et mutation :

La cellule sélectionnée à l'étape précédente (cellule mémoire ou cellule moyenne) est clonée proportionnellement à sa valeur de similarité. Le nombre de clones de chaque cellule est relatif à cette valeur, c'est-à-dire plus la similarité est grande plus on crée des clones. Tous les clones vont subir au processus de mutation. La mutation est réalisée de manière inverse à la valeur de similarité, c'est à dire plus grande est la similarité moins on mute les clones et vice versa. C'est un changement aléatoire d'une ou plusieurs valeurs des descripteurs du clone. Les processus de clonage et de mutation permettent de générer de nouvelles cellules qui peuvent être plus efficaces dans les prochaines générations.

4. Re-sélection des meilleurs clones par RBF :

La similarité entre les clones mutés et l'exemple d'apprentissage est calculée et comparée à celle de la cellule mémoire d'origine (sélectionné à l'étape 2 de l'apprentissage du SIA). Si la valeur de similarité du clone est plus faible que celle de la cellule mémoire, le clone est automatiquement rejeté. Tous les clones mutés restants sont présentés au réseau RBF correspondant à la même classe. Si un clone appartient à la même catégorie de sa cellule mémoire d'origine, il est ajouté à l'ensemble de cellules mémoires directement à la catégorie à laquelle il appartient, sinon il est rejeté. A la fin de l'apprentissage les différentes catégories de cellules mémoires obtenues pour chaque classe sont fusionnées pour être utilisés dans la classification. Le diagramme intégral de l'approche LDC-AIS est donné dans la figure 4.5.

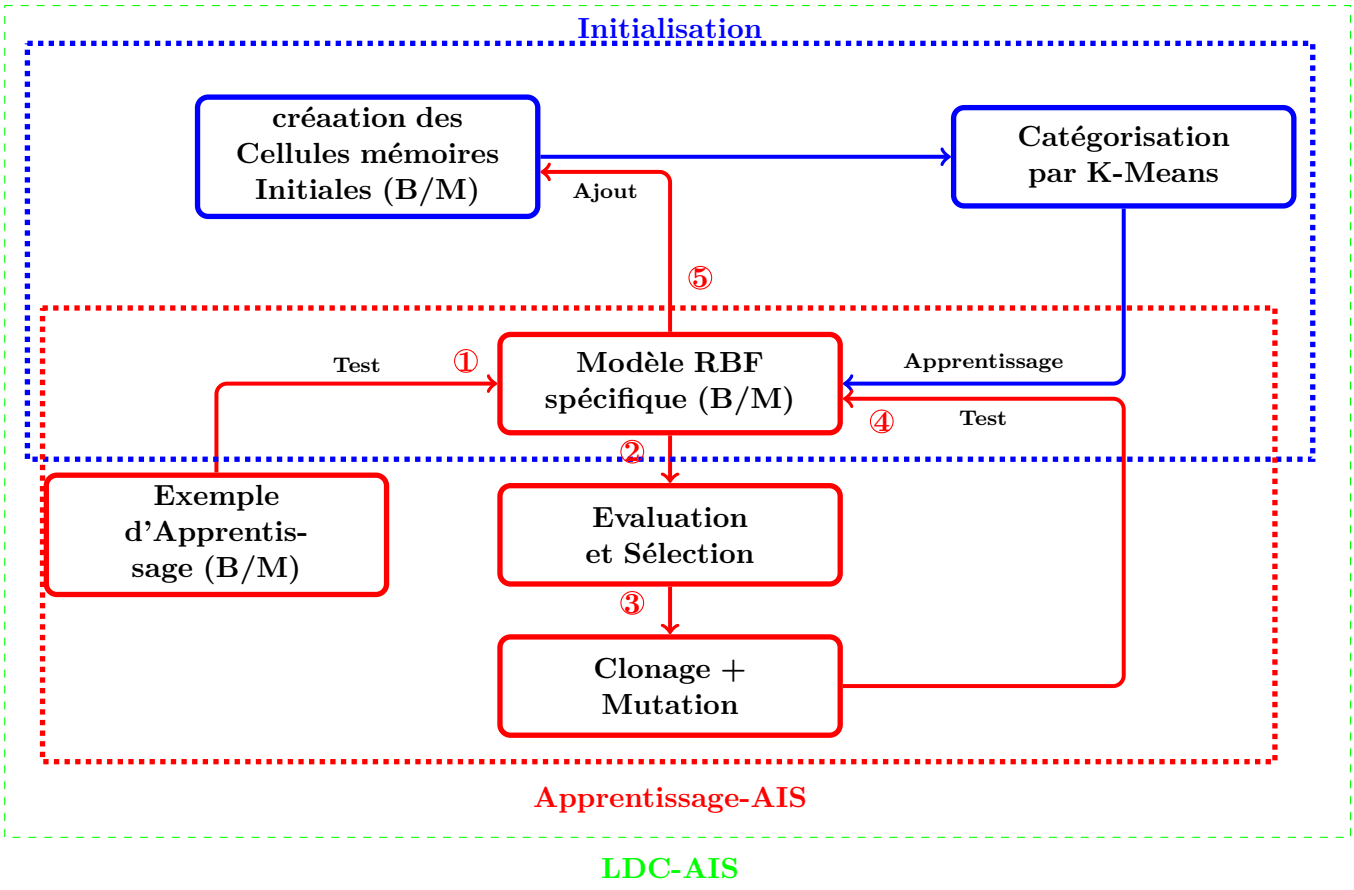


FIGURE 4.5 – Diagramme de LDC-AIS composé de l'étape d'initialisation (créatation des cellules mémoires initiales, Catégorisation par k-means, Apprentissage des catégories par RBF) et l'étape d'apprentissage du SIA (1.Test par RBF , 2.Evaluation et Sélection, 3.Clonage et Mutation, 4.Test des clones mutés par RBF et 5.Ajout des meilleurs clones aux cellules mémoires finales).

4.4 Expérimentations et résultats

Les résultats des différentes expérimentations effectuées sont présentés dans cette section. En premier lieu, on donnera les paramètres utilisés par chaque algorithme (K-Means, RBF et SIA). Ensuite, les résultats moyens de dix exécutions de LDC-AIS en termes de taux de classification et temps d'apprentissage seront détaillés sur chacune des deux bases de données WDBC et DDSM séparément. On finira par une comparaison des résultats avec d'autres approches immunitaires.

4.4.1 Paramètres utilisés

Avant de lancer un algorithme d'apprentissage, il est important de choisir les bons paramètres qui maximisent sa performance. Dans cette approche, chacun des algorithmes utilisé présente des paramètres à fixer : la valeur k (nombre de catégories) pour l'algorithme K-Means, la largeur de la gaussienne (largeur radiale du centre de base σ) du réseau RBF. Pour le SIA, le facteur de clonage (β), le nombre de clones (NB_{Clones}) ainsi que le nombre de cellules mémoires initiales NB_{CM_i}.

— Pour définir le nombre (k) de catégories de chaque classe d'apprentissage, nous avons appliqué une méthode simple et efficace qui consiste à catégoriser les exemples d'apprentissage (antigènes) par similarités. En effet, après la création des populations des cellules mémoires initiales, les similarités entre ces dernières et les antigènes de la même classe sont calculées et arrondies à un seule chiffre après la virgule. Le but c'est de créer des catégories de cellules mémoires spécifiques aux exemples d'apprentissage (antigènes). On peut ainsi avoir 11 catégories au maximum et une seule au minimum ($0 \leq Sim \leq 1$).

  Le nombre k de catégories dépend donc des données d'apprentissage et des cellules mémoires initiales, il est variable d'une classe à une autre et d'une base de données à une autre. Il peut aussi changer pour la même classe d'une base car différents exemples sont choisis pour l'apprentissage et le test dans chaque expérimentation. Les figures 4.6 et 4.7 montres des exemples du nombre de catégories de cellules mémoires pour les bases WDBC et DDSM respectivement.

— La largeur de la gaussienne du réseau de neurones RBF a été fixée expérimentalement à 0.5 pour un seuil d'erreur de 10^{-2}.

— Concernant l'apprentissage du système immunitaire artificiel, nous avons appliqué les mêmes paramètres que nous avons fixé dans le chapitre précédent(tableau 3.2) en vue de comparer les résultats par la suite.

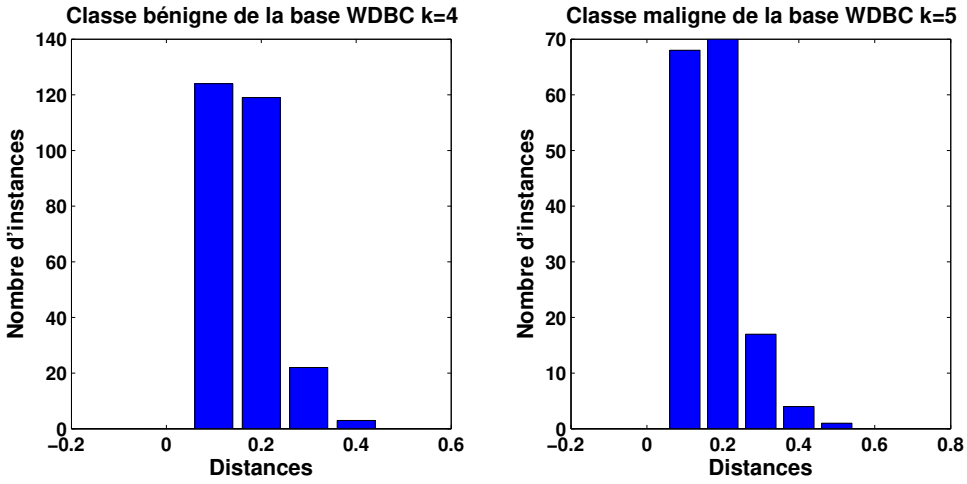


FIGURE 4.6 – Exemple d'histogrammes montrant les différentes catégories d'antigènes de la base WDBC (à gauche : classe bénigne 4 catégories, à droite : classe maligne 5 catégories

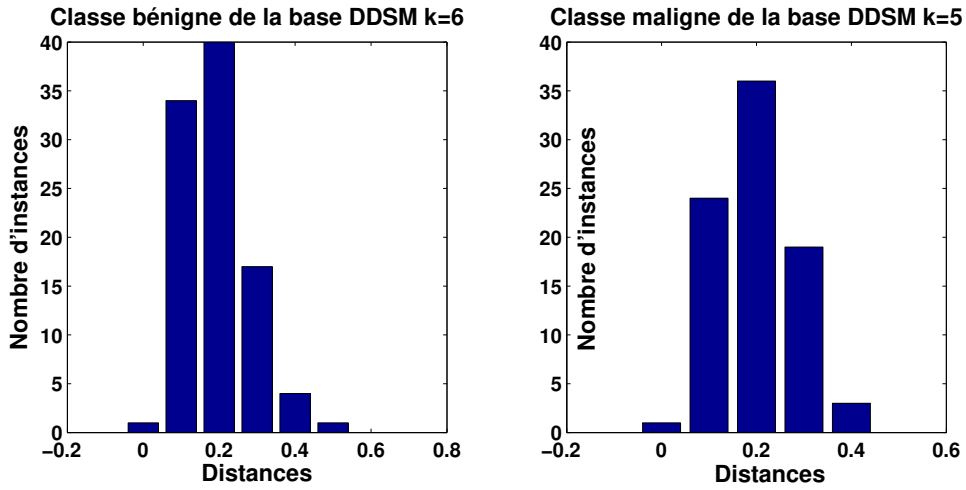


FIGURE 4.7 – Exemple d'histogrammes montrant les différentes catégories d'antigènes de la base DDSM (à gauche : classe bénigne 6 catégories, à droite : classe maligne 5 catégories

4.4.2 Résultats d'application sur la base WDBC

Cette section présentera les résultats d'application de LDC-AIS sur la base de données WDBC. Comme nous l'avons expliqué auparavant, nous avons appliqué notre algorithme avec deux différentes méthodes de sélection. La première consiste à sélectionner une seule cellule mémoire maximisant la similarité avec l'antigène pour le clonage et la mutation. La seconde méthode de sélection est celle que nous avons détaillé dans le chapitre précédent (approche AC-CLONALG). Le concept est de sélectionner P meilleurs cellules mémoires, pour en créer une cellule moyenne, qui sera en compétition avec la cellule la plus proche de l'antigène. La cellule la plus compétente entre les deux sera sélectionnée pour le clonage et la mutation. Nous avons appelé nos approches LDC-AIS(1) et LDC-AIS(2) respectivement pour les première et deuxième méthodes de sélection. Les résultats de chacune de ces deux méthodes sont présentés ci-dessous.

4.4.2.1 Résultats de la 1ère méthode de sélection

Le tableau 4.1 présente les résultats d'application de l'approche LDC-AIS sur la base WDBC en utilisant la première méthode de sélection.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)	Temps d'apprentissage (s)
2	95.36± 0.75	94.17± 1.42	42.23
5	95.61± 0.81	94.26± 0.96	145.54
10	95.70± 0.72	94.94± 0.49	441.91

Table 4.1 – Résultats de LDC-AIS sur la base WDBC avec la 1ère méthode de sélection

4.4.2.2 Résultats de la 2ème méthode de sélection

Dans le tableau 4.2, on présente les résultats de l'approche LDC-AIS sur la base WDBC en appliquant la seconde méthode de sélection (cellules moyennes).

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)	Temps d'apprentissage (s)
2	97.49± 0.69	94.78± 1.87	38.31
5	98.01± 0.58	95.82± 0.60	91.39
10	98.64± 0.50	97.52± 0.43	180.94

Table 4.2 – Résultats de LDC-AIS sur la base WDBC avec la 2ème méthode de sélection

La figure 4.8 illustre les courbes des résultats de temps d'apprentissage et taux de tests des deux méthodes LDC-AIS(1) et LDC-AIS(2) respectivement pour la première et deuxième méthode de sélection.

Sur la base WDBC, on remarque que les taux d'apprentissage et de test dépendent du nombre de générations. En effet, plus on augmente le nombre de génération de l'algorithme LDC-AIS (1 ou 2), plus les résultats sont meilleurs pour les deux méthodes de sélection.

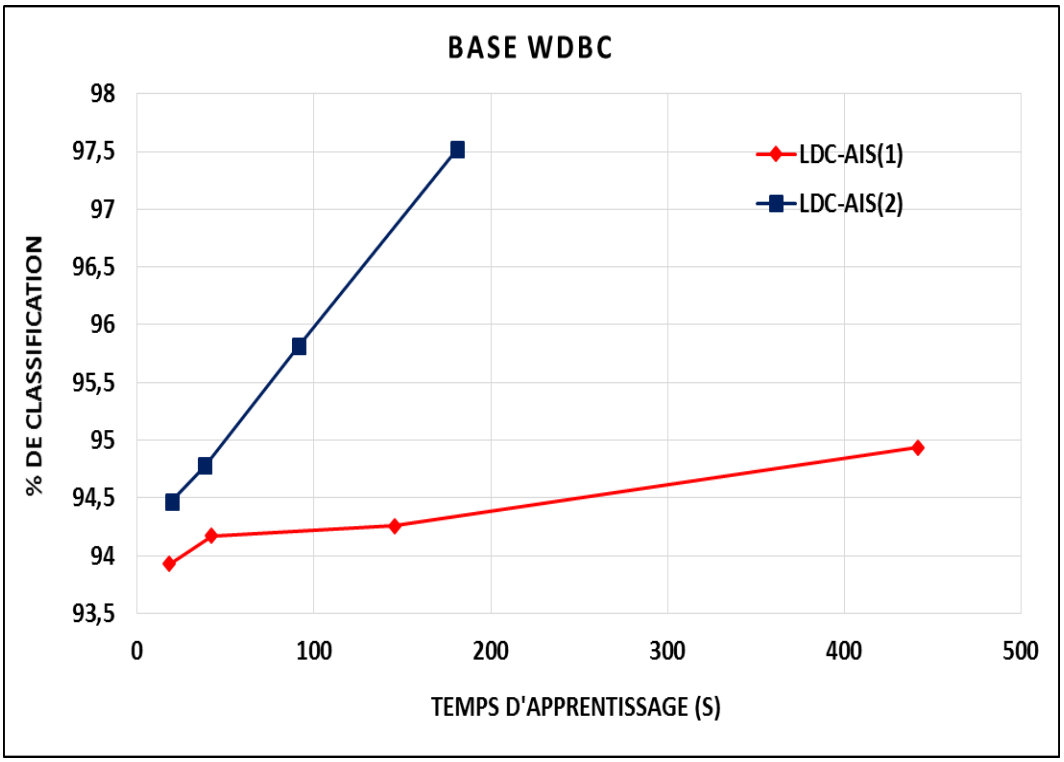


FIGURE 4.8 – Courbes des résultats de LDC-AIS(1) et LDC-AIS(2) sur WDBC

Par contre, il est facilement remarquable que la seconde méthode de sélection, qui est la sélection par création de cellules moyennes est largement meilleure que la première méthode. Les résultats de sélection d'une seule cellule pour le clonage sont moins bons.

Concernant le temps d'apprentissage, la différence entre les deux méthodes de sélection commence à s'élargir à partir de 5 générations. On remarque effectivement que l'amélioration des taux de classification est accompagnée par une accélération de l'apprentissage en utilisant la sélection par moyennes de cellules, contrairement à la méthode de sélection classique qui est plus lente.

La comparaison entre les deux courbes des résultats (figure 4.8) illustre l'efficacité de l'approche proposée en utilisant la création de cellules moyennes pour le clonage. On peut facilement conclure que cette méthode est bien plus rapide et efficace.

4.4.3 Résultats d'application sur la base DDSM

Comme sur la base de données précédente, nous avons appliqué notre algorithme sur la base DDSM en utilisant les deux méthodes de sélection. Les résultats des expérimentations effectuées incluant les taux d'apprentissage et de test ainsi que le temps d'apprentissage pour différents nombre d'itérations sont présentés dans les sections suivantes.

4.4.3.1 Résultats de la 1ère méthode de sélection

Le tableau 4.3 détaille les résultats expérimentaux de l'approche LDC-AIS sur la base de données DDSM avec la méthode de sélection d'une seule cellule pour le clonage (1ère méthode).

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)	Temps d'apprentissage (s)
2	94.35± 0.94	93.78± 0.63	63.48
5	94.74± 1.03	94.34± 0.85	166.80
10	94.91± 0.83	94.39± 0.24	347.54

TABLE 4.3 – Résultats de LDC-AIS sur la base DDSM avec la 1ère méthode de sélection

4.4.3.2 Résultats de la 2ème méthode de sélection

Les résultats d'application de l'algorithme LDC-AIS avec la méthode de sélection des cellules moyennes (seconde méthode) sont détaillés dans le tableau 4.4.

NB_{Gen}	Taux d'Apprentissage (%)	Taux de Test(%)	Temps d'apprentissage (s)
2	95.45± 1.15	94.43± 1.03	14.72
5	95.84± 0.71	94.96± 0.87	35.32
10	96.54± 0.68	95.62± 0.50	64.24

TABLE 4.4 – Résultats de LDC-AIS sur la base DDSM avec la 2ème méthode de sélection

Les courbes des deux méthodes de sélection sur la base DDSM sont illustrées dans La figure 4.9.

Sur la base de données DDSM, les remarques qui concernent l'amélioration des résultats de classification par l'augmentation du nombre de générations de l'algorithme sont identiques que ceux de la base précédente.

Si on compare les résultats entre les deux bases de données, on remarque que les taux de classification de la base WDBC sont meilleurs que ceux de la base DDSM. Cela est dû à consistance de cette dernière qui est inférieure. Par contre, la consistance de la base WDBC en nombre d'exemples et de descripteurs nécessite plus de temps de traitement. Cela explique pourquoi le temps d'apprentissage de la base WDBC est supérieur au temps d'apprentissage de la base de données DDSM.

Si on compare les deux méthodes de sélection, il est clair que la méthode de création de cellules moyennes pour le clonage est largement plus efficaces que ce soit en précision ou en rapidité. Cela confirme les résultats obtenus dans le chapitre précédent. Nous avons, effectivement, expliqué que la création des cellules moyennes spécifiques pour le clonage

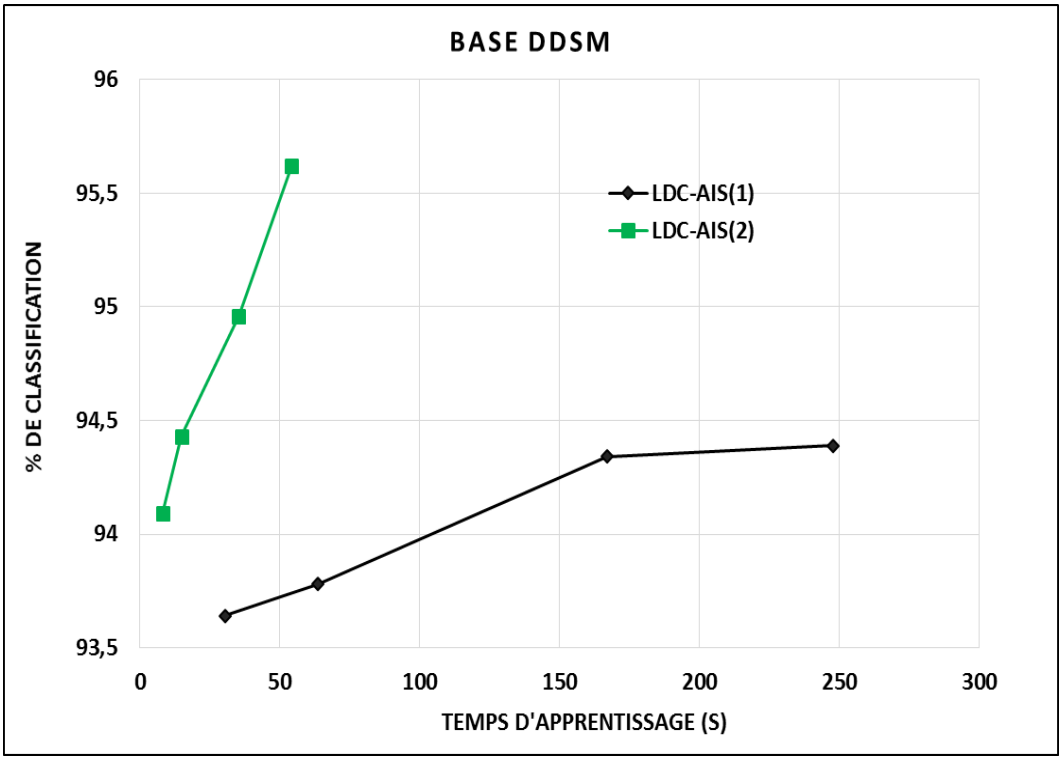


FIGURE 4.9 – courbes des résultats de LDC-AIS(1) et LDC-AIS(2) sur DDSM

améliore la représentativité des données d'apprentissage. De plus, dans l'algorithme LDC-AIS, la diversité est améliorée en sélectionnant seulement les clones qui appartiennent à la même catégorie que celle de l'exemple d'apprentissage.

La méthode LDC-AIS(1) sélectionne une seule cellule mémoire pour le clonage, cela veut dire que l'ensemble de ces clones mutés appartiennent probablement à la même catégorie. Ils vont donc tous rejoindre cette dernière. Tandis-que la seconde méthode LDC-AIS(2) crée des cellules potentielles, les clones de ces dernières n'appartiennent pas forcément à la même catégorie de l'exemple à apprendre. Par conséquent, la première méthode de sélection génère plus de cellules finales que la seconde.

Pour confirmer cette hypothèse, nous avons comparé les nombre de cellules mémoires finales de chaque méthode de sélection (LDC-AIS (1) et LDC-AIS(2)) par rapport au nombre de cellules mémoires initiales. Sur la base de données WDBC, le nombre de cellules initiales était de 129 cellules Bénignes (B) + 94 cellules Malignes(M). A la fin de 10 générations de LDC-AIS (1), le nombre de cellules mémoires finales était de 4559 cellules B + 1856 cellules M. Le nombre des cellules finales résultantes de la méthode de sélection LDC-AIS(2) était : 444 B + 293 M.

Puisque la base de données DDSM est moins consistante que la première, le nombre de cellules initiales ou finales est inférieur. En effet, à partir de 66 cellules B et 57 cellules M initiales, on est arrivé à 1337 + 1103 cellules bénignes et malignes respectivement en utilisant la première méthode de sélection. Le nombre des cellules mémoires finales obtenues par la seconde méthode était : 191 cellules bénignes + 134 cellules malignes.

Les taux de classification expliquent que ce n'est pas la quantité des cellules mémoires

qui importe mais leur qualité. En effet, même si nous avons proposé d'ajouter de nouvelles cellules à l'ensemble des cellules mémoires finales, il est important de bien choisir ces cellules.

La comparaison de ces résultats avec les approches de la littératures et celles présentées au chapitre précédent est faite dans la section suivante.

4.4.4 Comparaison des résultats

Dans cette section, on présente une étude comparative entre notre algorithme LDC-AIS et l'algorithme CLONALG ainsi que les approches proposées dans le chapitre précédent (MF-CLONALG, AC-CLONALG et VI-CS). Les tableaux 4.5 et 4.6 présentent les résultats de classification (%) et le temps d'apprentissage (s) de chacun de ces algorithmes sur les bases de données WDBC et DDSM respectivement. Les histogrammes de ces résultats sont illustrés dans les figures 4.10 et 4.11 respectivement.

Algorithme	Classification (%)	Temps d'apprentissage (s)
CLONALG	89,86	185,54
MF-CLONALG	95,03	374,48
AC-CLONALG	96,8	287,32
VI-CS	97,58	201,5
LDC-AIS(1)	94,94	441,91
LDC-AIS(2)	97,52	180,94

Table 4.5 – Comparaison des résultats WDBC

Algorithme	Classification (%)	Temps d'apprentissage (s)
CLONALG	85,15	62,88
MF-CLONALG	94,91	180,2
AC-CLONALG	94,98	178,66
VI-CS	95,76	154,12
LDC-AIS(1)	94,39	347,54
LDC-AIS(2)	95,62	64,24

Table 4.6 – Comparaison des résultats DDSM

A partir des tableaux de comparaison des résultats (4.5 et 4.6), on voit que les meilleurs taux de classification sont obtenus avec les méthodes VI-CS et LDC-AIS(2). Cependant, le temps nécessaire pour l'apprentissage est largement réduit dans la méthode LDC-AIS(2) comparé à toutes les versions améliorées de CLONALG, sur les deux bases de données.

Par exemple, sur la base WDBC, l'algorithme LDC-AIS(2) réduit le temps d'apprentissage de plus de 10% par rapport à VI-CS. De même, sur la base DDSM, il le réduit de plus de 58%. Par ailleurs, la méthode LDC-AIS(2) nécessite un temps d'apprentissage relativement similaire à celui de CLONALG, mais avec une nette amélioration du taux de classification.

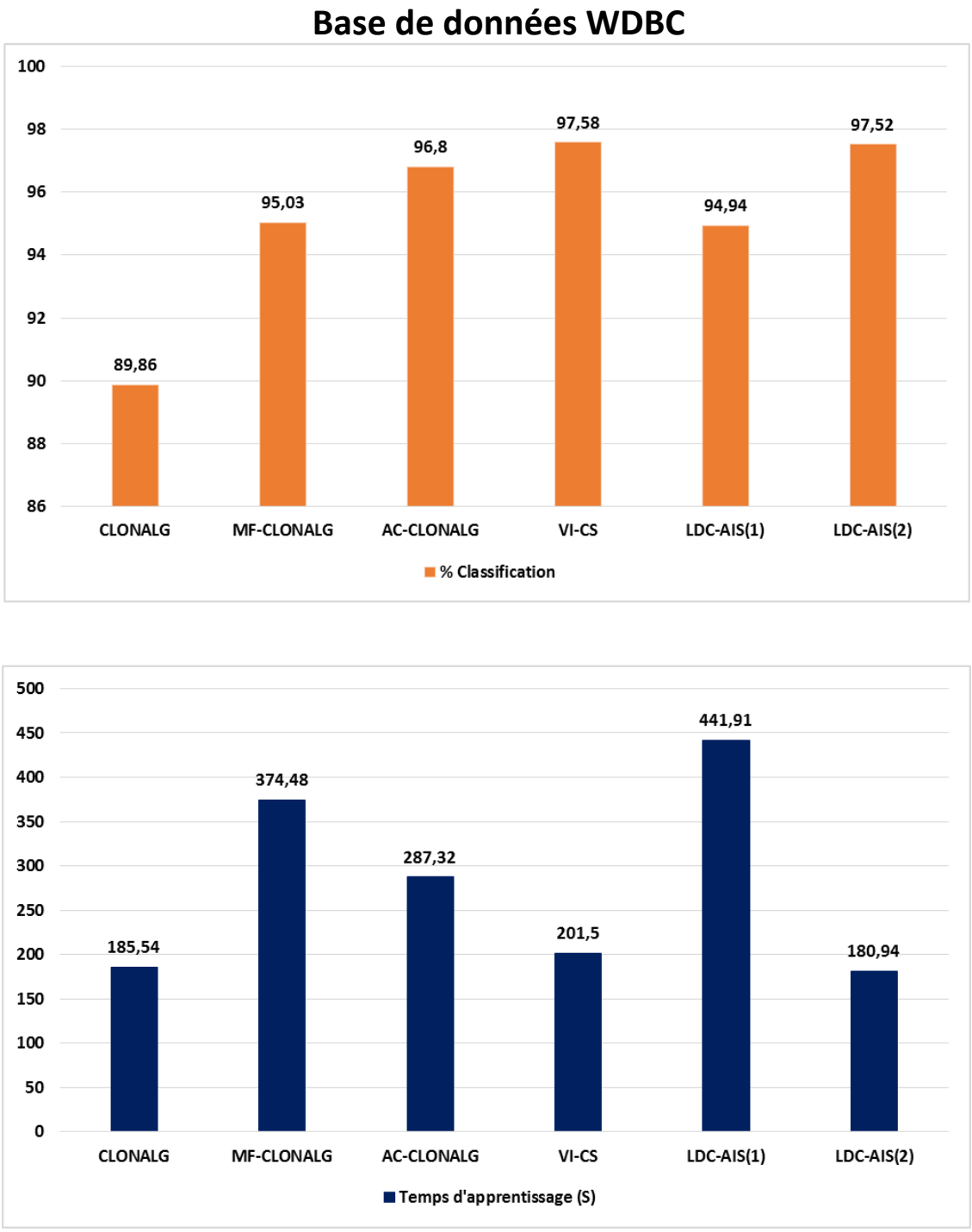


FIGURE 4.10 – Histogrammes de comparaison entre les taux de classification (haut) et le temps d'apprentissage (bas) des différentes algorithmes de sélection clonale sur la base de données WDBC

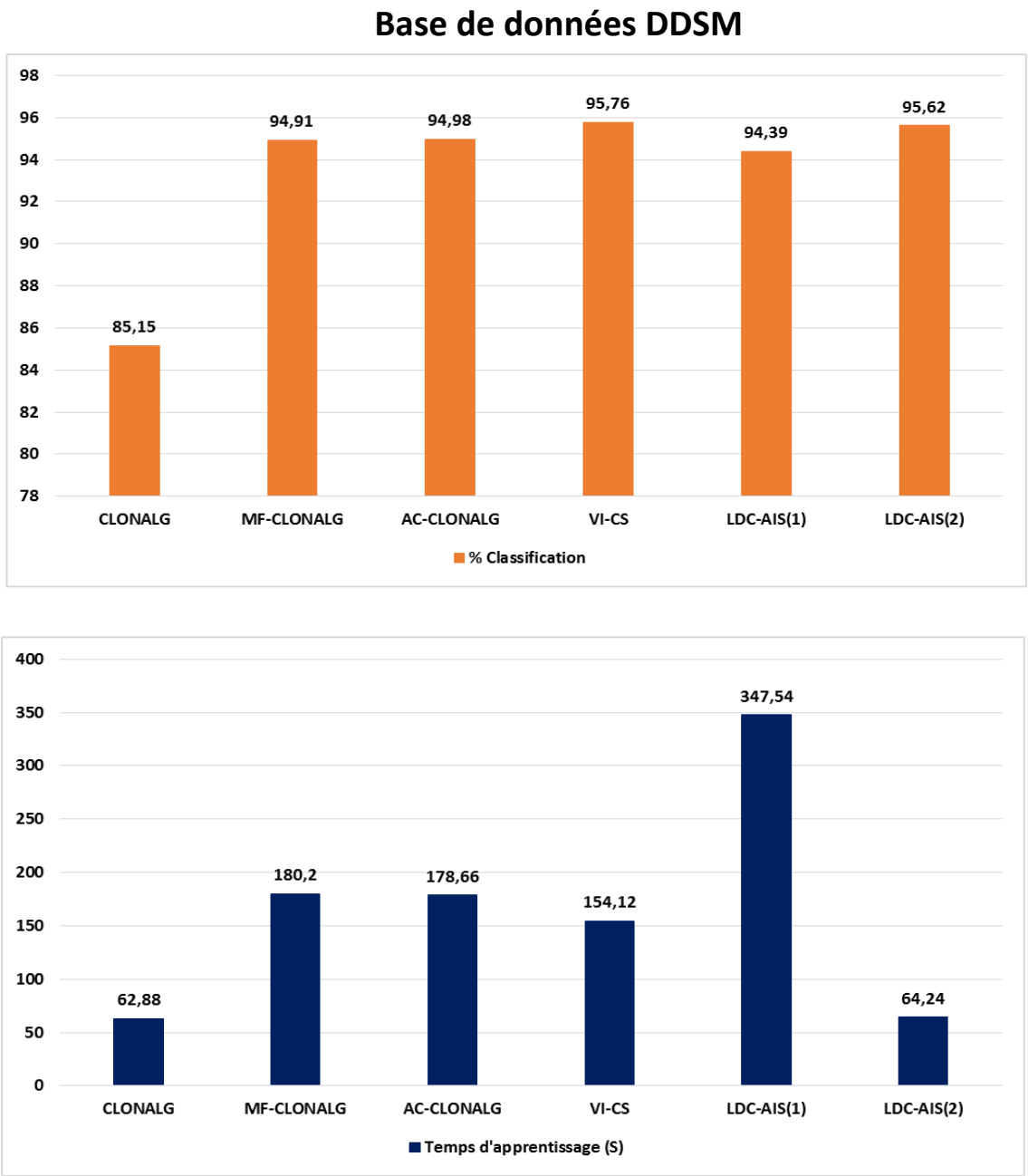


FIGURE 4.11 – Histogrammes de comparaison entre les taux de classification (haut) et le temps d'apprentissage (bas) des différentes algorithmes de sélection clonale sur la base de données DDSM

Ce qu'il faut retenir de ces expérimentations, c'est que la méthode LDC-AIS(2) présente à la fois les avantages de l'algorithme CLONALG en termes de rapidité, et ceux de VI-CS en termes de précision. La précision est garantie grâce aux cellules mémoires pertinentes ajoutées, et la rapidité au moyen de catégorisation locale de ces cellules.

4.5 Conclusion

Le premier objectif de ce chapitre était de trouver une solution pour le temps de calcul coûteux des algorithmes proposés dans le chapitre précédent en particulier, et ceux de la sélection clonale en général. Le second était celui d'éviter la perte de la performance de ces algorithmes. Dans ce but, nous avons présenté une approche avancée qui intègre l'algorithme K-Means et le réseau de neurones RBF à l'apprentissage des systèmes immunitaires artificiels. L'approche proposée utilise un mécanisme simple pour faire face à des fonctions de test coûteuses en temps. Les résultats de classification des bases de données WDBC et DDSM indiquent qu'un tel mécanisme, en dépit de sa simplicité est efficace et fournit de bonnes performances.

Chapitre 5

VERS UN PMC RAPIDE ET EFFICACE OPTIMISÉ PAR LA SLECTION CLONALE

5.1 Introduction

L'objectif principal de ce chapitre est d'introduire les Systèmes Immunitaires Artificiels (SIA) comme une technique d'optimisation bio-inspirée et d'étudier sa performance. En effet, après avoir présenté les algorithmes de sélection clonale comme un outil compétitif de reconnaissance des formes et de classification, nous nous sommes intéressés dans ce chapitre à explorer ce concept, afin de démontrer les avantages des opérateurs de clonage et de mutations dans le cadre de l'optimisation de fonctions.

La structure du chapitre est organisée comme suit : on commencera par définir l'optimisation et citer quelques travaux de la littérature utilisant la sélection clonale dans le traitement de ce problème. Par la suite, en vue d'exploiter le réseau de neurones Perceptron Multi-couches : PMC (Multi-Layer Perceptron : MLP en anglais) comme classifieur possédant une certaine performance prouvée théoriquement, nous avons eu l'idée de mettre l'optimisation par sélection clonale à l'épreuve. En réponse à certains inconvénients su PMC, un algorithme de réseau de neurones basé sélection clonale (MLP-CS) est mis en avant et discuté. Une procédure d'optimisation en plusieurs étapes est proposée, dans laquelle la rétropropagation est assistée par les processus de clonage et de mutation, pour une convergence plus rapide et précise du réseau PMC. La dernière partie du chapitre fournie les résultats d'application de MLP-CS et les compare avec ceux du PMC classique sur les bases de données WDBC et DDSM.

Enfin, comme l'approche MLP-CS est voisine des techniques d'optimisation évolutionnaires, une comparaison des résultats avec un PMC optimisé par un algorithme génétique en termes de précision et taux de calcul est accomplie, démontrant la performance des opérateurs des SIA.

5.2 Sélection Clonale Artificielle et Optimisation

Rechercher une solution optimale d'un problème constitue une préoccupation majeure dans différents domaines, que ce soit pour optimiser la sécurité, le coût, le confort, les gains ou le temps, etc.

En général, l'optimisation est la recherche des paramètres optimaux d'un système. Ces paramètres sont connus en tant que variables objectives [Kra14]. Le but de l'optimisation est de trouver les meilleurs éléments X^* possibles à partir d'un ensemble X, selon un ensemble de critères $F = f_1, f_2, ..., f_n$. Ces critères sont dits fonctions objectives [Wei09]. La résolution d'un problème d'optimisation vise donc à maximiser (ou minimiser) la fonction objective.

Ces dernières années, plusieurs travaux montrent l'émergence de nouvelles techniques d'optimisation, dont le principe se base sur la recherche de solutions en prenant en compte l'incertitude et l'imprécision de l'information réelle. Le but de ces techniques est celui de trouver des solutions satisfaisantes à coût approprié (au lieu de solution exactes) en utilisant l'apprentissage.

Une classe célèbre de ces techniques d'optimisation sont les stratégies évolutionnaires, qui ont prouvé leur succès dans les espaces à valeurs réelles. Ces méthodes permettent d'échapper des optimums locaux et de surmonter la stagnation prématurée.

Les techniques d'optimisation évolutionnaires regroupent divers algorithmes qui simulent les principes biologiques de l'évolution naturelle, tels que la compétition entre les individus, le croisement, la mutation, la sélection, la reproduction, etc. [Kra14][Ber10].

Dans ce chapitre, on s'intéresse aux Systèmes Immunitaires Artificiels (SIA), qui font partie de ces techniques évolutionnaires. En effet, parmi les différents algorithmes évolutionnaires d'optimisation tels que les algorithmes de colonies de fourmis (Ant Systems), les algorithmes génétiques (Genetic Algorithms), ou les algorithmes à essaims de particules (Partical Swarm Optimisers), on trouve les SIA qui se sont saisi de l'optimisation comme domaine prometteur d'application, conduisant à un certain nombre de publications relatives aux problèmes d'optimisation de fonctions. La majorité de ces publications sont basées sur l'application du principe de sélection clonale, déclarant souvent un certain succès comparé aux autres algorithmes de la littérature.

Dans [DCVZ02], les auteurs ont présenté CLONALG pour effectuer principalement l'apprentissage et les taches de reconnaissance de formes. Ensuite, il a été adapté pour effectuer des tâches d'optimisation en introduisant les modification suivantes :

1. Au lieu de la population d'antigènes (AG) à reconnaître (classifier), on trouve la fonction objectif à optimiser (maximiser ou minimiser).
2. La mesure de similarité correspond à l'évaluation de la fonction objective pour une cellule mémoire donnée.
3. Le clonage n'est plus proportionnel à la similarité, car toutes les cellules ont un nombre identique de clones. Par contre, la mutation reste toujours inversement proportionnelle à la valeur de similarité (valeur d'évaluation de la fonction objective) de la cellule mémoire.

Les résultats d'application sur différents problèmes ont montré que la seconde version

de CLONALG est une technique appropriée pour l'optimisation de fonctions, surtout quand il s'agit d'optimisation multimodale où l'espace du problème est constitué de plusieurs solution optimales locales.

Un autre algorithme immunitaire basé sur la sélection clonale a été proposé dans [SGHL06] pour l'optimisation multimodale. Dans ce travail, les auteur modifient le principe de maintenance de la diversité des cellules mémoires en calculant la concentration de chacune de ces dernières grâce à sa similarité moyenne. Les résultats de simulation ont montré que l'algorithme obtient efficacement la solution optimale tout en maintenant une bonne diversité. Comparé à CLONALG pour l'optimisation, et opt-aiNet (Artificial Immune Network for Optimization) [dCT02a], l'algorithme d'optimisation modifié présente une performance supérieure pour localiser l'optimum global dans les espaces de grandes dimensions.

Basé sur la sélection clonale artificielle et l'apprentissage Baldwinien (qui consiste à guider le processus d'évolution grâce à l'apprentissage) [Bal96] [Tur97] [Fed03], les auteurs de [GJZ10] ont proposé un algorithme d'optimisation nommé Baldwinian Clonal Selection Algorithm (BCSA). C'est le 1er travail qui a introduit l'apprentissage baldwinien au SIA. L'approche utilise un nouvel opérateur d'apprentissage baldwinien pour simuler l'effet baldwinien dans la sélection clonale artificielle. L'algorithme BCSA évolue et améliore la population des cellules mémoires en appliquant cet opérateur baldwinien aux clones avant de les muter. Les résultats expérimentaux de BCSA sur différents types de fonctions, ainsi que sur un problème réel, ont montré que l'algorithme réussi à accomplir avec succès la résolution des problèmes d'optimisation.

Le problème de couverture par ensembles (set covering problem) [Ste06] est un problème d'optimisation combinatoire [Sak84]. Le premier travail arrivant à résoudre ce problème en utilisant les systèmes immunitaires artificiels a été publié dans [TRR12]. Les auteurs utilisent un CLONALG modifié pour introduire systématiquement l'aléatoire dans l'algorithme, afin de construire une solution faisable. La modification de CLONALG a été dans les étapes de sélection et suppression des individus. L'algorithme AIS-SCP remplace les cellules mémoires moins bonnes par les meilleurs clones mutés. A la fin de l'apprentissage, les exemples aléatoires ne sont introduits à l'ensemble des cellules mémoires seulement s'ils ont une meilleure similarité que les cellules qu'ils remplacent. Les résultats d'application de AIS-SCP ont prouvé l'efficacité de ce dernier dans la génération de solutions de haute qualité du problème de couverture par ensembles.

Les auteurs de [PL15] proposent un algorithme de sélection clonale hybride pour l'apprentissage et l'optimisation (Hybrid Learning Clonal Selection Algorithm). HLCSA utilise deux mécanismes d'apprentissage pour guider le processus de réponse immunitaire : l'apprentissage baldwinien [Tur97] et l'apprentissage orthogonal [ACC00]. Après l'opération de clonage dans l'algorithme immunitaire hybride, l'apprentissage baldwinien est exécuté comme dans [GJZ10] pour aider à améliorer les informations contenues dans la population de cellules mémoires, en modifiant l'espace de recherche. L'apprentissage orthogonal vient ensuite pour remplacer l'étape de mutation des clones et compléter l'apprentissage baldwinien. Les résultats obtenus sur divers problèmes ont prouvé l'efficacité de l'algorithme HLCSA dans l'optimisation.

Tous les travaux cités ci-dessus prouvent la performance des algorithmes des SIA,

et spécifiquement ceux de la sélection clonale dans différentes optimisations. Grâce aux opérateurs de clonage et mutation, l'espace de recherche est mieux exploré pour aider à trouver efficacement l'optimum global. Ces raisons nous ont poussés à étudier la performance de la sélection clonale artificielle dans le domaine d'optimisation de fonctions.

Dans ce contexte, le réseau de neurones Perceptron Multi-Couches (PMC) (en anglais MLP : Multi-Layer Perceptron) est le modèle des réseaux de neurones le plus utilisés dans la littérature pour la classification du cancer du sein. Néanmoins, ce dernier présente certains inconvénients qui sont la lente convergence et la stagnation dans les minima locaux. En tant que tel, et pour répondre à ces inconvénients, nous proposons d'explorer le concept de la sélection clonale afin de démontrer les avantages des opérateurs de clonage et de mutations dans le cadre de l'optimisation des poids du PMC pour une convergence rapide vers le minimum global.

Pour toutes ces motivations, on utilise dans ce chapitre les principes de la sélection clonale pour optimiser le réseau de neurone Perceptron Multi-Couches. On présentera dans la section qui suit le principe général du PMC et on discutera son inconvénient qui a besoin d'être optimisé et les différents travaux effectués pour cette raison.

5.3 Perceptron Multi-Couches (PMC)

Les Réseaux de Neurones Artificiels (RNA) représentent une technique d'apprentissage qui a connu une évolution constante depuis les travaux de Mac Culloch et Pitts en 1943 [MP43]. Ce sont des modèles informatiques qui emploient des formes simplifiées du système neuronal biologique, offrant une excellente résolution des problèmes dans de nombreux domaines. Les neurones du réseau sont interconnectés grâce à des liens de connexion dont chacun possède un poids (w), ces poids sont multipliés par le signal transmis dans le réseau. La sortie du réseau est déterminée en utilisant une fonction d'activation telle que la sigmoïde ou la gaussienne [KSJ+00] [AJMA07]. Les réseaux de neurones à rétropropagation d'erreur (en anglais Back-Propagation Neural Networks BBNNs) sont un type de base de réseaux de neurones basé sur la méthode de descente du gradient (Gradient Descent (GD)) [Sny05], qui tente de réduire au minimum l'erreur du réseau en abaissant le gradient de la courbe d'erreur.

Une classe importante des BPNNs est le Perceptron Multi-Couches (PMC) (Multi-Layer Perceptron Neural Network (MLP) en anglais). Ce modèle d'apprentissage supervisé est composé d'une couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie. Les connexions entre les nœuds (neurones) des couches adjacentes transmettent les signaux de sortie d'une couche à la couche suivante. Les principales caractéristiques du PMC sont : le fonctionnement rapide, la facilité d'implémentation, la capacité d'apprentissage et de généralisation, voilà pourquoi ils sont les architectures de réseaux de neurones les plus couramment utilisées [BH00]. Un exemple d'un réseau de neurones PMC à une seule couche cachée est fourni dans la figure 5.1.

Bien que le PMC est adapté à une grande variété d'applications, il présente néanmoins une lenteur de la convergence. En effet, à cause de l'initialisation aléatoire des poids, et

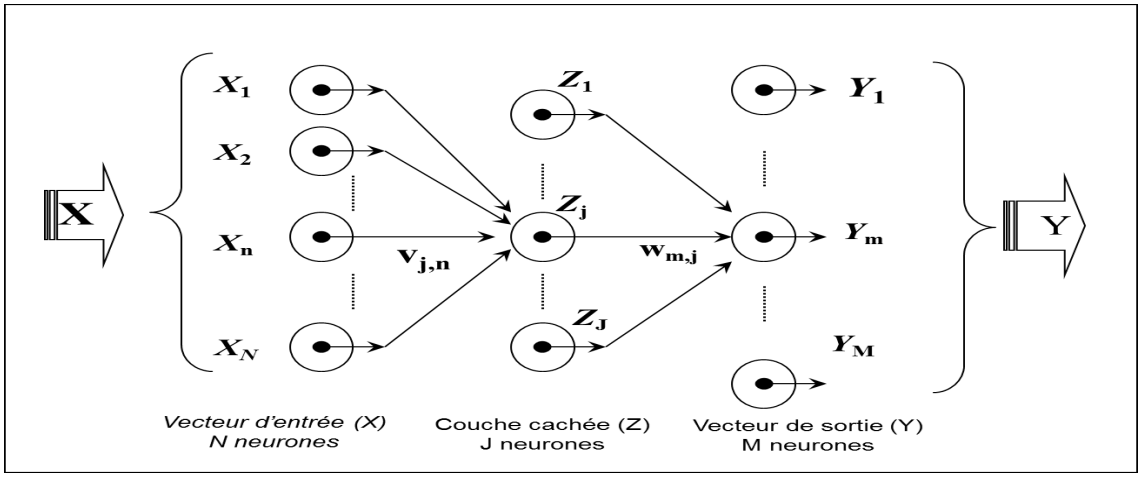


FIGURE 5.1 – Exemple de réseau de neurones PMC avec une seule couche cachée.

le temps nécessaire à l'algorithme de descente du gradient pour l'adaptation du poids du réseau, le PMC converge lentement vers l'optimum global [RHW85] [SJW92] [RB93] [AKV08].

Diverses améliorations ont donc été apportées à ce dernier afin d'accélérer sa convergence. Dans [VMR+88], les auteurs ont appliqué des modifications à l'algorithme de rétropropagation décrit dans [RHW85] afin d'accélérer la convergence. Les modifications impliquent la mise à jour des poids seulement après la présentation de tous les exemples d'apprentissage au réseau, au lieu de les mettre après la présentation de chaque exemple. Une autre modification consiste à faire varier dynamiquement le pas d'apprentissage (LR pour Learning Rate en anglais). Le travail proposé dans [HNGS11] consiste à modifier l'algorithme de rétropropagation du gradient par l'introduction du gain adaptatif en même temps qu'un pas d'apprentissage adaptatif. Les simulations informatiques ont montré que ces modifications ont donné un meilleur taux de convergence par rapport à la rétropropagation classique. Un algorithme de rétropropagation avec un pas d'apprentissage et une inertie (momentum) dynamiques (Back Propagation with Dynamic training Rate and Momentum (BPDRM)) a été publié dans [AYYA15]. L'algorithme propose une nouvelle stratégie qui consiste en plusieurs étapes, pour éviter l'inflation dans les poids, en ajoutant chaque pas d'apprentissage et momentum comme fonctions dynamiques. Les expériences ont montré que BPDRM fourni un apprentissage plus rapide par rapport à l'algorithme de rétropropagation classique au même seuil d'erreur.

Bien que les modification citées ci-dessus soient efficaces, elle ne traitent que le problème de la rapidité de convergence sans se soucier de la précision du réseau de neurones. En outre, la majorité des améliorations ont été apportées à l'algorithme de rétropropagation, et non au PMC. Ainsi, il demeure important de trouver une approche qui accélère la convergence des PMC, tout en améliorant ses performances.

Dans ce contexte, diverses méthodes d'optimisation ont été développées, souvent basées sur des approches évolutionnaires, car ces méthodes ont l'avantage de combiner les règles et l'aspect aléatoire. Parmi ces algorithmes on trouve les algorithmes génétiques

(AG). Les AG constituent un outil d'optimisation globale puissant, néanmoins ils présentent quelques limitations dans la recherche locale. En dépit de leurs résultats pertinents, les AGs nécessitent un temps considérable pour converger vers une solution optimale. En effet, le caractère aléatoire du croisement signifie que les AG ne parviennent pas à produire de bons individus, ce qui augmente le nombre d'itérations nécessaires pour produire les solutions recherchées. D'autres approches évolutionnaires ont ensuite été développées.

Étant donné que le système immunitaire humain fonctionne comme un mécanisme cognitif qui reconnaît des modèles qui peuvent ne pas être reconnus par le système nerveux, le SIA peut être utilisé en tant que complément aux architectures des réseaux de neurones. Par ailleurs, bien que le comportement du système immunitaire est très complexe, nous avons plus d'informations sur ce dernier que sur le système nerveux. Par conséquent, de nouvelles structures basées sur l'immunologie peuvent être proposées comme alternative aux RNA.

Nous avons vu dans la section 5.2 que les SIA ont réussi à prouver leur efficacité dans différentes optimisations. Dans ce chapitre, nous proposons d'étudier la performance de la sélection clonale dans l'optimisation du Perceptron Multi-Couches en approfondissant la recherche de poids optimaux du réseau. Dans l'approche que l'on propose, les opérateurs de clonage et de mutation sont utilisés pour permettre une convergence rapide et une meilleure précision du PMC, sans aucune modification de l'algorithme de rétropropagation du gradient. En effet, dans le PMC, les poids du réseau peuvent être considérés comme des cellules mémoires qu'il est nécessaire d'adapter aux antigènes (les exemples d'apprentissage) afin de minimiser l'erreur de sortie du réseau (fonction objectif).

5.4 Perceptron Multi-Couches basé Sélection Clonale

Le but principal d'un réseau de neurones est de mettre à jour les poids de connexions afin de minimiser la distance entre les sorties obtenues et les sorties désirées. Cette distance est représentée par l'erreur quadratique moyenne (EQM) entre les deux sorties, calculée par l'équation suivante :

$$EQM = \frac{\sum_{k=1}^{N}(D_k - Y_k)^2}{N} \tag{5.1}$$

où N est le nombre total des exemples d'apprentissage, D_k est la sortie désirée du k$^{\text{ème}}$ exemple, et Y_k, la sortie obtenue du k$^{\text{ème}}$ exemple. La minimisation de l'EQM implique l'obtention des poids optimaux du réseau.

Afin d'aider le PMC à trouver les poids optimaux qui minimisent l'erreur quadratique moyenne plus rapidement et efficacement en même temps, le principe de l'approche proposée consiste à employer les avantages des systèmes immunitaires artificiels, par l'application des opérateurs de clonage et de mutation pendant l'apprentissage du PMC. L'idée de MLP-CS (Multi-Layer Perceptron based Clonal Selection) est de cloner les poids mis à jour par la rétropropagation classique, et de muter chaque clone afin d'obtenir des poids plus efficaces. Ce processus permettra au PMC de converger rapidement vers le seuil d'erreur quadratique moyenne (EQM_{th}) défini par l'utilisateur.

5.4.1 Mise à jour des poids du PMC par rétropropagation

La rétropropagation est l'outil le plus utilisé dans les réseaux de neurones artificiels. C'est une technique de calcul des dérivées basée sur l'algorithme de descente du gradient [Sny05]. Elle peut être appliquée sur n'importe quelle structure de fonctions dérivables. Dans cette méthode, l'erreur de sortie du réseau est propagée vers les couches cachées, d'où le nom rétropropagation [HN89]. Dans le PMC, le principe de rétropropagation est utilisé pour calculer la variation de poids adéquate ($\Delta W_{j,i}$) afin de les mettre à jour. La mise à jour des poids du réseau est effectuée à la fin d'apprentissage de chaque exemple de la base selon l'équation :

$$W_{j,i+1} = W_{j,i} + \Delta W_{j,i} \tag{5.2}$$

Dans un PMC classique avec une fonction sigmoïde, la révision des poids se fait de manière récursive depuis le sortie vers l'entrée du réseau à partir de la couche de sortie. Soit e_j l'erreur calculée pour le neurone j obtenue par :

$$e_j = D_j - Y_j \tag{5.3}$$

où D_j correspond à la sortie désirée et Y_j la sortie obtenue du neurone j calculée par :

$$Y_j = f(\sum_i W_{j,i}.X_i) \tag{5.4}$$

X_i correspond à l'entrée du neurone j. Dans notre travail, on considère la partie principale de l'optimisation PMC comme suit :

— $f(x)$ est la fonction sigmoïde définie par :

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.5}$$

L'objectif est d'adapter les poids du réseau de manière à minimiser l'erreur quadratique moyenne (EQM). Chaque poids est modifié d'une valeur de $\Delta W_{j,i}$ où :

$$\Delta W_{j,i} = \alpha.Y_i.\delta_j \tag{5.6}$$

Avec :

— $W_{j,i}$: les poids de connexion entre deux neurones j et i de deux couches successives ;

— α : le pas d'apprentissage (LR) $0 < \alpha < 1$;

et

$$\delta_j = \begin{cases} e_j.Y_j.(1 - Y_j) & si \quad j \text{ est un neurone de la couche de sortie} \\ Y_j.(1 - Y_j).\sum_k \delta_k.W_{k,j} & if \quad j \text{ est un neurone de la couche cachée} \end{cases} \tag{5.7}$$

avec k est le nombre de neurones dans la couche cachée. La procédure détaillée de la mise à jour des poids par l'algorithme de rétropropagation du gradient est expliquée dans Annexe C.

Après la rétropropagation, nous obtenons les poids mis à jour $[W_i]$ ayant une EQM inférieure que celle des poids avant la mise à jour $[W_{i-1}]$. Notre idée est d'utiliser le principe

de la sélection clonale des systèmes immunitaires artificiels, pour aider la rétropropagation et accélérer la recherche des poids optimaux, par le clonage et la mutation de ces poids mis à jour. Ces opérateurs permettent la production de nouveaux poids qui pourraient présenter une erreur quadratique inférieure à celle obtenue en utilisant la rétropropagation. Le schéma de l'approche proposée est illustré à la Figure 5.2.

5.4.2 Sélection clonale des meilleurs poids du PMC

L'objectif principal du système immunitaire est de reconnaître les cellules étrangères (antigènes) qui attaquent le corps. Si la différence entre l'anticorps (ou cellule mémoire) et l'antigène est faible, la similarité entre cet antigène et l'anticorps est grande, ce qui signifie que la reconnaissance est plus probable. Dans notre travail, la similarité est elle-même l'Erreur Quadratique Moyenne (EQM). Les antigènes sont les exemples d'apprentissage de la la base de données, et les cellules mémoires sont représentées par les poids du réseau PMC. Le but est alors de minimiser la similarité (EQM) entre les antigènes et les cellules mémoires. A cet effet, et comme mentionné précédemment, nous utiliserons les opérateurs de clonage et de mutation des SIA pour aider le PMC à trouver plus rapidement les poids optimaux, qui offrent une erreur inférieure au seuil EQM_{th}.

Le mot clonage se réfère à l'acte d'isoler un objet ou une personne, et de le multiplier à l'identique. Dans les SIA, le clonage consiste à créer des copies identique de la (les) cellule(s) mémoire(s) sélectionnée(s) afin d'améliorer sa (leur) compétitivité. Dans notre algorithme, à la fin du processus de clonage, on obtient N copies identiques des poids mis à jour par rétropropagation $[W_i]$ à l'étape précédente (section 5.4.1). L'ensemble des clones est : $[W_i^1.....W_i^N]$, où N représente le nombre de clones. Le deuxième opérateur important dans les algorithme SIA est la mutation. Elle consiste en un changement aléatoire d'une ou plusieurs positions des valeurs caractéristique du clone. Ce processus est destiné à former les clones sélectionnés pour devenir des cellules mémoires plus ou moins bonnes, les moins bons clones sont éliminés par la sélection, il ne restera que les meilleurs.

Dans notre algorithme MLP-CS , la mutation consiste à réviser les valeurs des clones par un processus aléatoire dans le but de générer une meilleure matrice de poids que la matrice originale, à savoir des poids présentant une plus petite erreur (EQM_{Min}) que celle obtenue par rétropropagation.

Après l'étape de mutation, on procède à l'évaluation des clones mutés : $[W_m^1.....W_m^N]$ par le calcul de la similarité de chacun, et on sélectionne le meilleur clone : $[W_m^k]$ qui maximise la valeur de similarité ($EQM_{BestClone}$), avec $k \in [1, \ldots, N]$, et m indiquant les clones mutés.

Cette valeur de similarité ($EQM_{BestClone}$) est comparée à celle de la cellule originale : $[W_i]$ (EQM_{BP}), ensuite une mise à jour des poids $[W_{i+1}]$ est effectuée en remplaçant les poids $[W_i]$ avec ceux qui présentent une l'EQM minimale (EQM_{Min}).

L'EQM de la nouvelle cellule $[W_{i+1}]$ est comparée au seuil d'erreur EQM_{th} fixé par l'utilisateur au début de l'algorithme, si cette valeur est supérieure au seuil, on retourne à l'étape de rétropropagation pour créer de nouvelles cellules d'une nouvelle itération, jusqu'à l'obtention d'une cellule avec une EQM inférieure à EQM_{th}.

Cette cellule finale représente la solution de l'algorithme, c'est-à-dire, les poids optimaux du PMC qui seront utilisés pour la classification du cancer du sein. L'algorithme 1

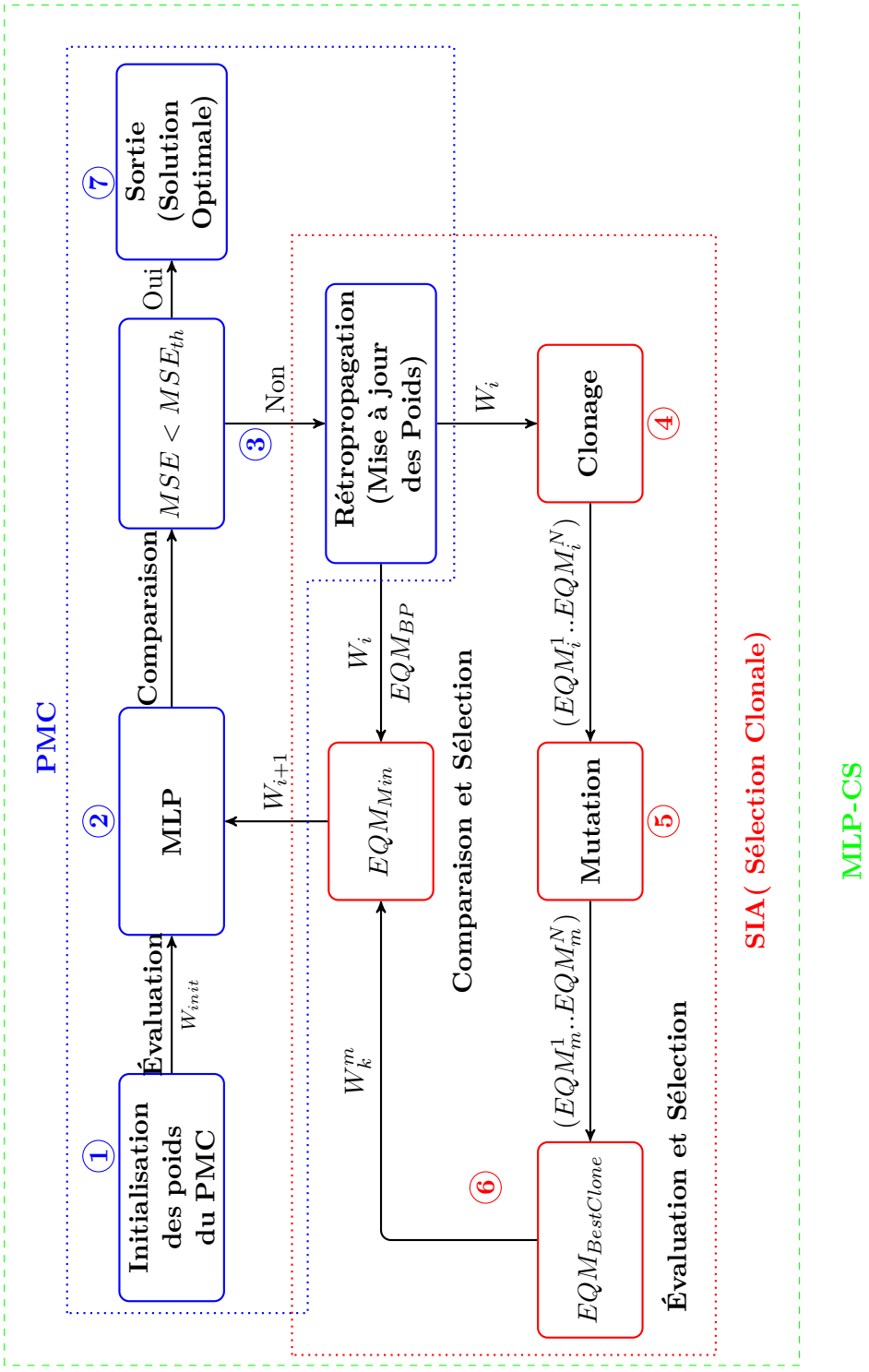


FIGURE 5.2 – Diagramme de l'approche MLP-CS composé de : 1. Initialisation des poids du PMC, 2. Évaluation (calcul de EQM_{BP}), 3. Rétropropagation et mise à jour des poids, 4. Clonage des poids mis à jour, 5. Mutation des clones, 6. Évaluation des clones mutés, comparaison avec les poids mis à jour par rétropropagation et sélection des meilleurs poids avec EQM_{Min}, 7. Comparaison entre EQM et EQM_{th} et obtention des poids optimaux du PMC minimisant l'EQM.

fournit la description algorithmique de l'approche MLP-CS proposée.

La Figure 5.3 explique les étapes de clonage et de mutation suivis par la sélection des poids optimaux produisant EQM_{Min} .

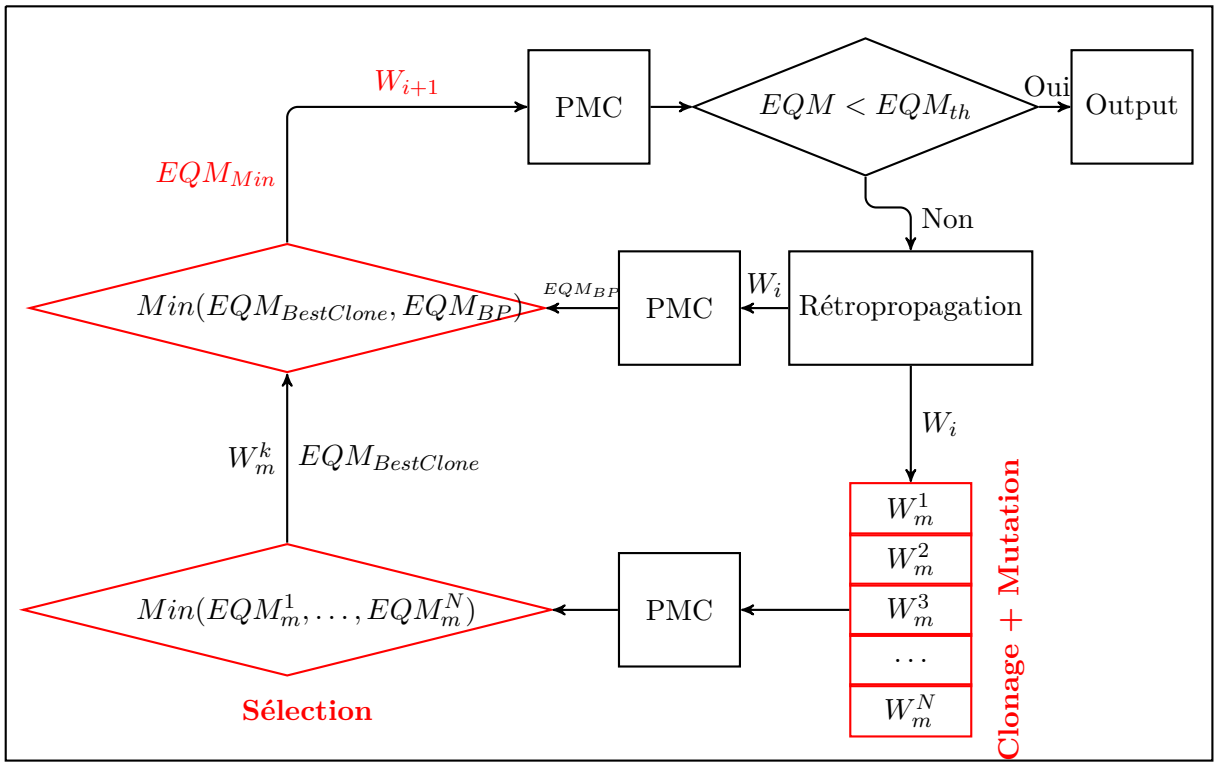


FIGURE 5.3 – Après la rétropropagation, les poids mis à jour sont clonés et mutés séparément, et le meilleur clone muté $[W_m^k]$ ayant l'EQM minimale est sélectionné pour être comparé aux poids mis à jour par rétropropagation. Les poids produisant MSE_{Min} sont sélectionnés pour être les nouveaux poids d'une nouvelle itération et sont comparés à MSE_{th}

.

5.5 Résultats Expérimentaux

Cette section regroupera les résultats d'application de l'algorithme MLP-CS sur les deux bases de données WDBC et DDSM pour la classification du cancer du sein. On présentera les paramètres utilisés et les résultats obtenus sur chaque base séparément. on commencera par la base WDBC, en suite la base DDSM, et on finira par une comparaison des résultats.

5.5.1 Résultats de MLP-CS sur la base WDBC

Avant de présenter les résultats d'application de l'algorithme MLP-CS sur la base WDBC, on commercera par donner les différents paramètres de l'algorithme.

Algorithm 1 Pseudo-code de l'algorithme MLP-CS

1. Générer aléatoirement les poids initiaux du PMC : $[W_{init}]$
2. Évaluation (Calculer EQM)

while $EQM < EQM_{th}$ **do**

3. Rétropropagation (mise à jour des poids : $[W_i]$) et calcul de EQM_{BP}
4. Clonage des nouveaux poids : $[W_i^1.....W_i^N]$
5. Mutation des clones : $[W_m^1.....W_m^N]$
6. Réévaluation et sélection su meilleur clone muté : $[W_m^k]$

if $EQM_{BestClone}$ $([W_m^k]) < EQM_{BP}([W_i])$ **then**

Poids $[W_{i+1}]$ = Meilleur Clone : $[W_m^k]$

$EQM = EQM_{BestClone}$

else

Poids $[W_{i+1}] = [W_i]$

$EQM = EQM_{BP}$

end if

end while(itération)

5.5.1.1 Paramètres utilisés

Sur la base WDBC, le modèle PMC utilisé est composé d'une couche d'entrée à 30 neurones représentant les descripteurs de WDBC, une seule couche cachée où le nombre de neurones est déterminé expérimentalement, et une couche de sortie. La fonction de transfert est la sigmoïde pour les deux couches (cachée et de sortie).

Il est nécessaire de contrôler trois paramètres importants du PMC : le seuil d'erreur quadratique moyenne (EQM_{th}), le pas d'apprentissage (α) et le nombre de neurones dans la couche cachée (NNHL) . Pour déterminer la meilleure combinaison de ces trois paramètres, nous avons sélectionné cinq valeurs de EQM_{th} : de 10^{-1} à 10^{-5}, et nous avons appliqué un PMC en utilisant cinq valeurs de α : 0.1, 0.25, 0.5, 0.75 et 1, et cinq autres valeurs de : 5, 10, 12, 15 et 20 neurones.

Les résultats de classification (%) de chaque expérimentation sont listés dans les tableaux 5.1 et 5.2.

α \ EQM_{th}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.1	36.48	96.95	**99.31**	**99.31**	**99.31**
0.25	36.48	96.62	98.94	99.30	99.30
0.5	36.48	95.94	98.30	98.93	98.93
0.75	36.48	94.93	97.96	98.95	98.95
1	36.48	89.18	97.28	98.32	98.32

TABLE 5.1 – Résultats de classification (%) de différentes valeurs du seuil d'Erreur Quadratique Moyenne (EQM_{th}) et du pas d'apprentissage (α) sur la base WDBC.

A partir des tableaux 5.1 et 5.2, on peut remarquer que les meilleurs résultats sont obtenus avec α=0.1, 10 neurones dans la couche cachée et un seuil EQM_{th} de 10^{-3}.

NNHL \ EQM_{th}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
5	52.70	97.29	98.64	99.30	**99.31**
10	36.48	96.95	**99.31**	**99.31**	**99.31**
12	36.48	97.63	98.59	98.60	98.60
15	36.48	97.29	98.94	98.60	98.60
20	36.48	96.95	98.94	98.93	98.93

TABLE 5.2 – Résultats de Classification (%) de différentes valeurs du seuil d'Erreur Quadratique Moyenne ($EQMth$) et du nombre de neurones dans la couche cachée Layer (NNHL) sur la base WDBC.

Les résultats obtenus avec les seuils d'EQM de 10^{-4} et 10^{-5} sont presque les mêmes, mais ils nécessitent plus de temps pour converger. Par exemple, la convergence vers $EQM_{th} = 10^{-3}$ a pris 29.46 seconds tandis-qu'il a fallu 113 secondes au PMC pour converger vers le seuil $EQM_{th} = 10^{-4}$, et plus de 243 secondes pour converger vers $EQM_{th} = 10^{-5}$, alors que la précision de classification est la même : 99,31%. Le seuil d'erreur MSE_{th} est donc fixé à 10^{-3}.

Le tableau 5.3 résume l'ensemble des paramètres utilisés dans notre travail sur la base de données WDBC.

Paramètre	Valeur
Nombre d'entrées	30
Nombre de couches cachées	1
EQM_{th}	10^{-3}
α	0.1
NNHL	10
Nombre de clones (SIA)	5
Similarité	EQM
Fonction d'activation	Sigmoïde

TABLE 5.3 – Paramètre de l'algorithme MLP-CS utilisés sur la base WDBC.

5.5.1.2 Évaluation de MLP-CS et performances obtenues

Basé sur tests effectués pour définir les paramètres du PMC, nous avons utilisé dans ce travail la validation croisée avec k=5 (5-fold cross validation) pour évaluer la performance de l'approche proposée (MLP-CS).

Les résultats de chaque exécution des deux algorithmes (MLP (PMC) et MLP-CS) incluant les précision de classification (%), le temps de convergence (s) et le nombre total des itération sur la base WDBC sont donnés dans le tableau 5.4. Les résultats moyens des cinq exécutions sont illustrés dans la figure 5.4 et les courbes de convergence des deux algorithmes sont tracées dans la figure 5.5 .

Notons que la meilleure précision de classification est obtenue par notre approche MLP-CS qui a obtenu 99,02 % par rapport au PMC qui a obtenu 98,44 %. Le temps de

Algorithme	MLP					MLP-CS				
$N°Execution$	1^{ere}	2^{eme}	3^{eme}	4^{eme}	5^{eme}	1^{ere}	2^{eme}	3^{eme}	4^{eme}	5^{eme}
(%) classification	97.87	98.91	98.21	98.96	98.29	98.22	99.65	98.66	99.30	99.29
Temps d'exécution(s)	23.69	30.49	27.28	19.68	19.85	15.76	25.62	14.68	10.88	12.99
$N°$ Itérations	1113	1446	1289	943	955	373	615	347	258	313

TABLE 5.4 – Résultats de classification de 5 exécutions de MLP(PMC) et MLP-CS sur la base WDBC (précision, temps de convergence et nombre total d'itérations)

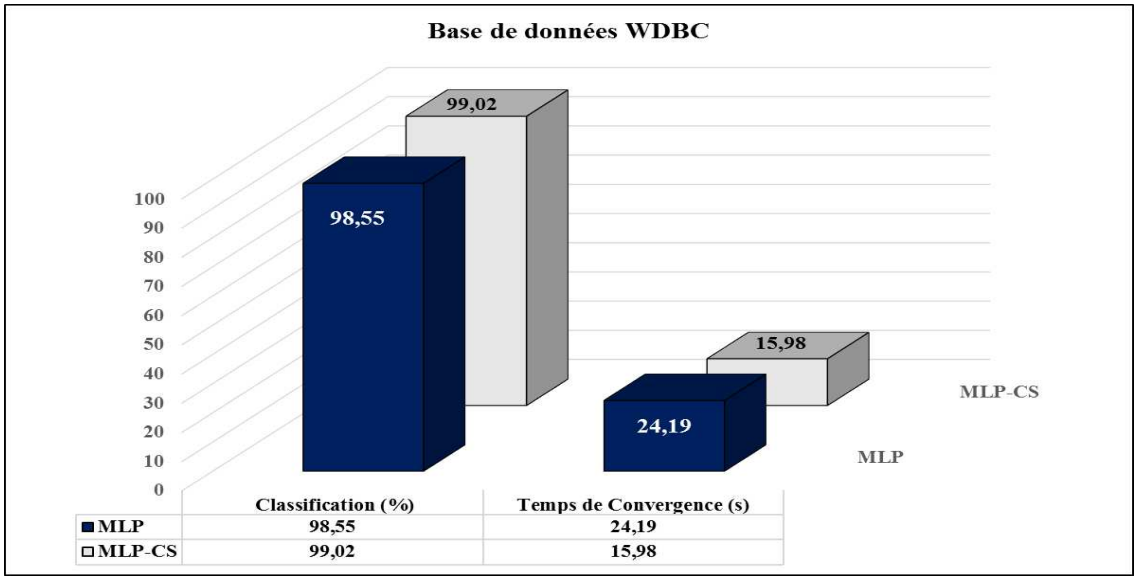


FIGURE 5.4 – Résultats moyens de classification et de temps de calcul de 5 exécutions de MLP et MLP-CS sur la base de données WDBC.

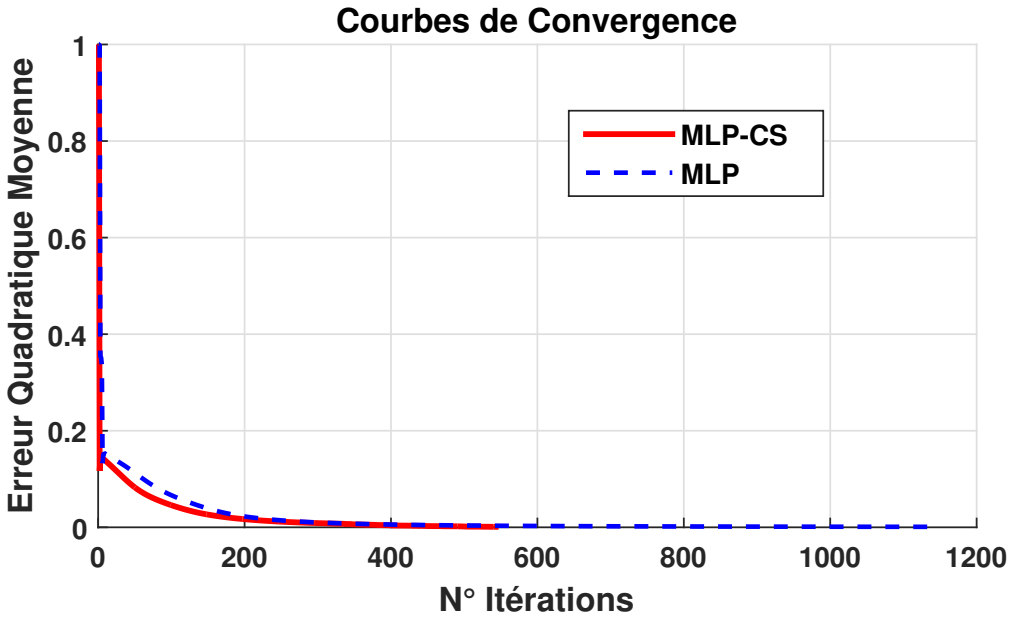


FIGURE 5.5 – Courbes de convergence de l'EQM en appliquant MLP et MLP-CS sur la base WDBC : MLP nécessite a plusieurs itérations pour converger vers EQM_{th} comparé à MLP-CS qui nécessite beaucoup moins d'itérations.

calcul est amélioré de 34 %. En effet, MLP-CS a pris une moyenne de 15.98 secondes pour converger vers le seuil d'erreur lorsque la convergence du PMC vers la même erreur a pris 24.19 secondes.

5.5.2 Résultats de MLP-CS sur la base DDSM

Après avoir présenté les résultats sur la base WDBC, on donnera dans cette section les résultats obtenus sur la base de données DDSM. Comme sur la précédente base, on commencera par la fixation des paramètres de l'algorithme, puis, on détaillera les résultats obtenus en utilisant ces paramètres.

5.5.2.1 Paramètres utilisés

Sur la base de données de DDSM, le modèle PMC utilisé est composé d'une couche d'entrée avec 22 neurones représentant les 22 descripteurs de la base, une seule couche cachée et une couche de sortie. Nous avons procédé de la même manière que pour la base de données WDBC pour définir les bons paramètres de PMC pour la base de données DDSM : EQM_{th}, α et NNHL sont fixés expérimentalement.

Les tableaux 5.5 et 5.6 listent les résultats de classification (%) des différents tests effectués.

A partir des tableaux 5.5 and 5.6, nous avons fixé le nombre de neurones dans la couche cachée (NNHL) à 12 neurones, le pas d'apprentissage α à 0.5 et le seuil d'erreur EQM_{th} à 10^{-3}. comme sur la base WDBC, les résultats utilisant les seuils 10^{-4} et 10^{-5} sont pratiquement les mêmes mais nécessitent plus de temps pour converger.

EQM_{th} / α	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
0.1	66.10	90.67	93.56	93.22	92.37
0.25	65.25	93.22	92.37	**94.06**	93.56
0.5	44.91	93.22	**94.06**	**94.06**	**94.06**
0.75	44.91	91.52	93.22	92.37	93.89
1	44.94	92.37	93.22	92.37	93.89

TABLE 5.5 – Résultats de classification (%) de différentes valeurs du seuil d'Erreur Quadratique Moyenne (EQM_{th}) et du pas d'apprentissage (α) sur la base DDSM.

EQM_{th} / NNHL	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
5	44.91	91.37	92.37	92.22	93.22
10	51.69	92.37	93.63	90.67	93.94
12	44.91	92.37	**94.06**	**94.06**	**94.06**
15	44.91	92.37	92.72	93.22	**94.06**
20	44.91	93.22	93.22	90.67	93.22

TABLE 5.6 – Résultats de classification (%) de différentes valeurs du seuil d'Erreur Quadratique Moyenne ($EQMth$) et du nombre de neurones dans la couche cachée (NNHL) sur la base DDSM.

L'approche proposée MLP-CS sur la base de données DDSM est évaluée en utilisant les paramètres listés dans le tableau 5.7.

Parametère	Valeur
Nombre d'entrées	22
Nombre de couches cachées	1
EQM_{th}	10^{-3}
α	0.5
NNHL	12
Nombre de clones	5
Similarité	EQM
Fonction d'activation	Sigmoïde

TABLE 5.7 – Paramètre de l'algorithme MLP-CS utilisés sur la base DDSM.

5.5.2.2 Évaluation de MLP-CS et performances obtenues

Après avoir fixé les paramètre de MLP-CS sur la base DDSM, nous avons procédé comme sur la base précédente, c'est à dire la validation croisée avec k=5. Les résultats de chacune des cinq exécutions des algotithmes MLP et MLP-CS comprenant le taux de classification (%), le temps d'exécution et le nombre d'itérations sur DDSM sont donnés dans le tablea 5.8.

Algorithme	MLP					MLP-CS				
$N°Execution$	1^{ere}	2^{eme}	3^{eme}	4^{eme}	5^{eme}	1^{ere}	2^{eme}	3^{eme}	4^{eme}	5^{eme}
(%) classification	91.80	94.91	94.91	92.37	94.06	94.26	95.76	96.61	95.41	95.76
Temps d'exécution(s)	31.52	12.18	29.42	13.47	23.44	29.67	10.95	19.39	10.54	15.87
$N°$ Itérations	3672	1261	2758	1173	2130	2321	835	1068	575	855

TABLE 5.8 – Résultats de classification de 5 exécutions de MLP(PMC) et MLP-CS sur la base de données DDSM (précision, temps de convergence et nombre total d'itérations)

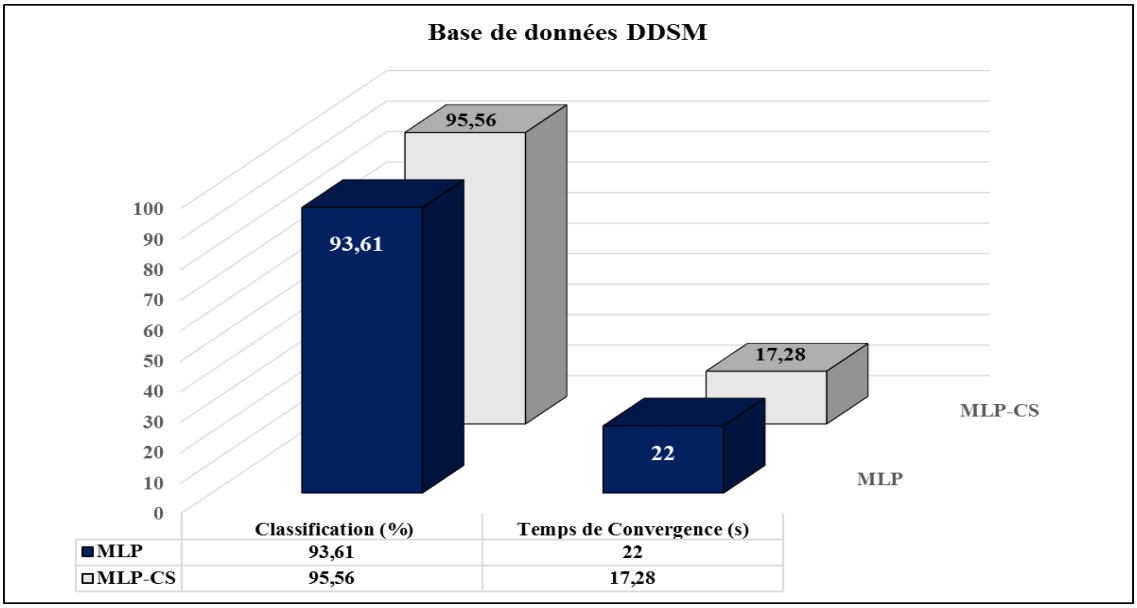


FIGURE 5.6 – Résultats moyens de classification et de temps de calcul de 5 exécutions de MLP et MLP-CS sur la base DDSM.

Après cinq exécutions, on remarque à partir de la figure 5.6, qui illustre les moyennes des résultats, que la précision du PMC est améliorée de 1.95%, et son temps de convergence est réduit par 21.5 %. En effet, MLP-CS a obtenu une moyenne de 95.56% de bonne classification en 17.28 secondes comparé au MLP qui n'a convergé qu'après 22 secondes pour obtenir un taux moyen de 93.61 %. Les courbes de convergence des deux algorithmes (PMC (MLP) et MLP-CS) sur la base de données DDSM sont tracées dans la figure 5.7.

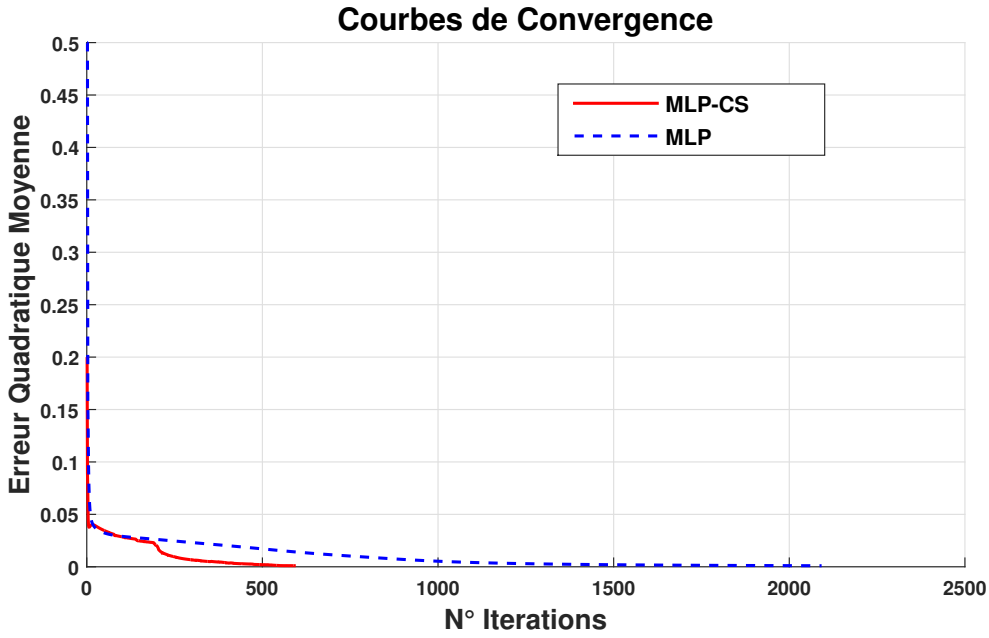


FIGURE 5.7 – Courbes de convergence de l'EQM en appliquant MLP et MLP-CS sur la base DDSM : MLP a pris 2800 itérations pour converger vers EQM_{th} alors que MLP-CS n'a pris que 550 itérations pour converger vers le même seuil d'erreur.

5.5.3 Discussion

Les résultats obtenus montrent clairement que l'amélioration apportée au PMC est efficace. En effet, l'approche proposée (MLP-CS) permet de réduire considérablement le temps d'apprentissage du réseau de neurones, et améliore légèrement son taux de classification. Sur la base de données WDBC, on constate une réduction de 34% du temps de convergence, accompagnée d'une amélioration de 0,58% de la précision de la classification. Sur la base de données DDSM, la réduction moyenne du temps de calcul est de 21,5%, et la précision est améliorée de presque 2%. On remarque aussi une grande réduction du nombre d'itérations nécessaires à la convergence vers EQM_{th}, et ce, malgré l'ajout des étapes de clonage et de mutation à l'algorithme d'apprentissage du PMC.

Nous présentons à travers la figure 5.8 une étude représentative sur la base de données WDBC, où l'EQM obtenue par rétropropagation est tracée en gris, et les différentes valeurs de l'EQM des clones mutés de la même matrice de poids sont tracées en noir. Nous pouvons voir que l'EQM_{Min} est celle obtenue par le meilleur clone. Les poids sont donc remplacés par ce meilleur clone muté dans cette itération.

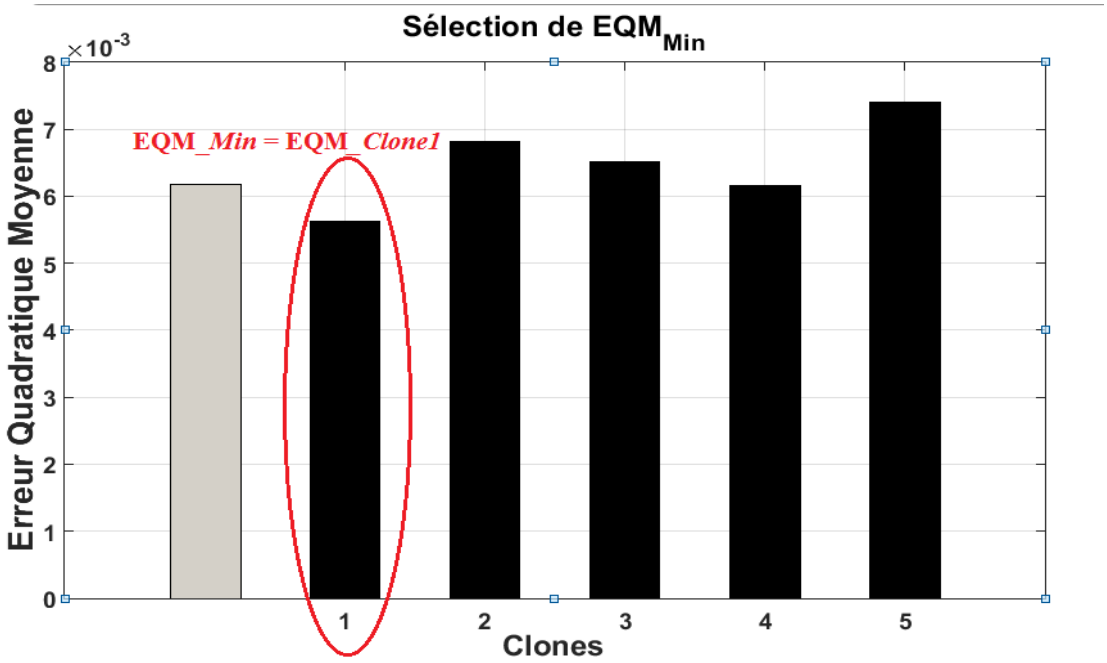


FIGURE 5.8 – Étude de cas : Différentes valeurs de EQM obtenues par rétropropagation (en gris) et par les clones mutés de (en noir) et sélection de EQM_{Min} par $Clone_1$ en utilisant l'opérateur des SIA : $[W_{i+1}] = Clone_1$.

Avec cette méthode, nous pouvons garantir une convergence plus rapide du PMC vers l'erreur quadratique moyenne définie par l'utilisateur. Les Figures 5.9 et 5.10 tracent les courbes de convergence de MLP-CS sur les bases de données WDBC et DDSM respectivement, et les histogrammes de sélection, c-à-d, chaque barre de l'histogramme représente une itération dans laquelle EQM_{Min} est obtenue grâce aux clonage et mutation (SIA). Cela signifie que $(EQM_{BestClone} < EQM_{BP})$. Comme on peut le voir, dans un grand nombre d'itérations, la convergence du PMC est accélérée en utilisant le principe de sélection clonale.

La figure 5.11 montre les différences entres ces valeurs d'EQM pour chaque itération : $(EQM_{BestClone} - EQM_{BP})$, et toutes ces valeurs sont négatives, ce qui prouve l'efficacité de l'approche MLP-CS.

Comme nous l'avons mentionné dans la section 5.2, parmi les méthodes évolutionnaires d'optimisation, on retrouve les algorithmes génétiques (AG) qui constituent un outil puissant ayant été largement utilisé dans diverses optimisations, y compris la classification du cancer du sein. En outre, les principes sont similaires à ceux des SIA. Dans ce contexte, et à des fins de comparaison, Tableau 5.9 présente une évaluation entre le MLP, MLP-CS et MLP-AG, que nous avons implémenté en utilisant les mêmes paramètres cités dans les tableaux 5.3 et 5.7 (le SIA a été remplacé par un algorithme génétique pour comparer les résultats), sur les deux bases WDBC et DDSM. Les résultats sont obtenus en utilisant un facteur 5 de la validation croisée, et ils montrent que, bien que l'AG améliore la précision et le temps de convergence du MLP, le SIA présente les meilleurs taux. Ces résultats

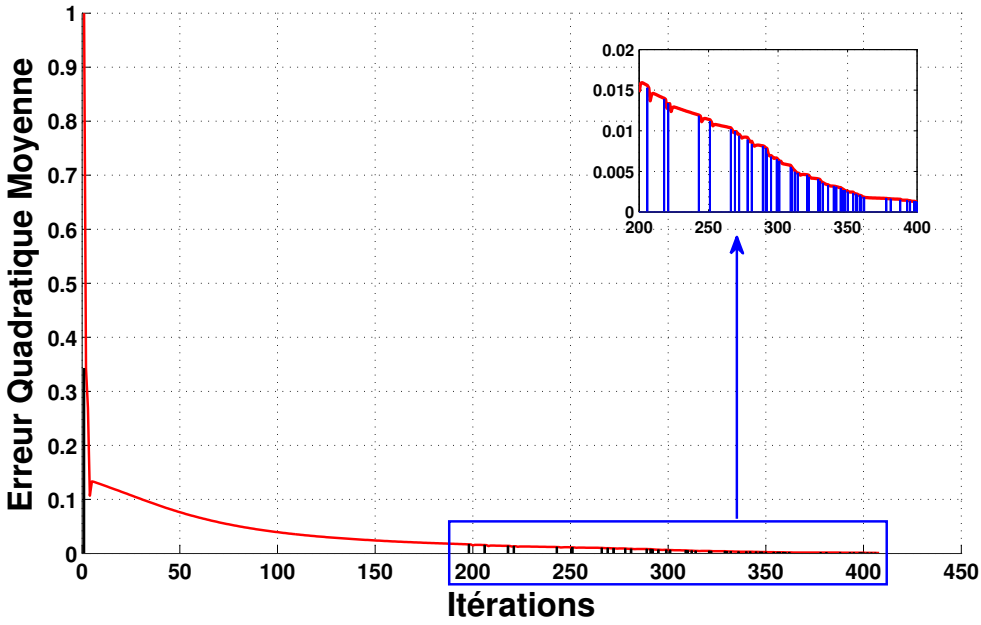


FIGURE 5.9 – Courbe de convergence de MLP-CS sur la base WDBC et histogramme de sélection : chaque barre de l'histogramme signifie que EQM_{Min} sélectionnée est obtenue par le clonage et la mutation permettant la réduction du nombre d'itérations nécessaires pour converger, ce qui prouve l'efficacité des opérateurs du SIA sur le PMC.

prouvent que l'opérateur de croisement de l'AG exige plus de temps pour converger vers le seuil d'erreur que l'opérateur de clonage de SIA qui est plus rapide et plus efficace.

Base de données	WDBC		DDSM	
Algorithme	Temps (s)	Classification(%)	Temps (s)	Classification (%)
MLP	24.19	98.44	22	93.61
MLP-CS	15.98	99.02	17.28	95.56
MLP-AG	21.19	98.53	19.5	94.07

TABLE 5.9 – comparaison entre l'algorithme génétique et MLP-CS

Pour finir, on présente dans le tableau 5.10 les résultats de classification de notre travail (MLP-CS) et ceux des études antérieures de la littérature appliquées sur la base de données WDBC. Comme on peut le constater à partir des résultats, l'approche MLP-CS proposée obtient une très bonne précision de classification.

5.6 Conclusion

Nous avons étudié dans ce chapitre la performance des systèmes immunitaires artificiels dans le cadre de l'optimisation. Notre objectif était d'optimiser le temps de convergence du Perceptron Multi-Couches (PMC) sans aucun effet sur sa performance de classifica-

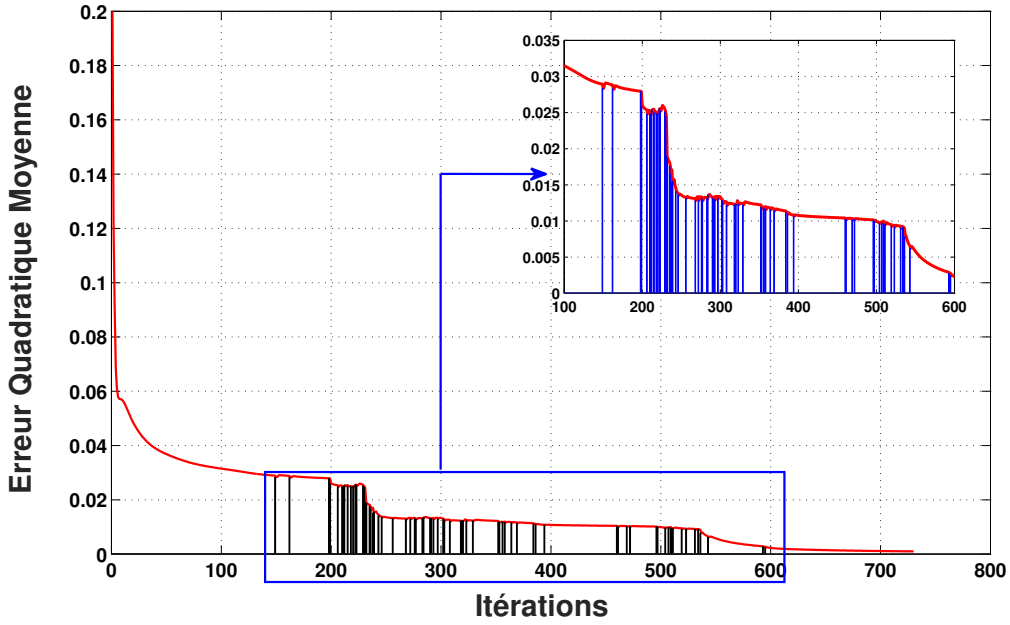


FIGURE 5.10 – Courbe de convergence de MLP-CS sur la base DDSM et histogramme de sélection : chaque barre de l'histogramme signifie que EQM_{Min} a été sélectionnée en utilisant le clonage et la mutation des poids au lieu de la rétropropagation, ce qui permet de réduire le nombre d'itérations nécessaires au PMC pour converger au seuil EQM_{th}.

Auteur	Méthode	Classification (%)
[GN06]	SVM	96.32%
[VSKT10]	Modular Neural Network	98.22%
	Radial Basis Function NN	97.63%
[SS11]	CLONALG	71.90%
	CLONAX	93.40
[BBÖ13]	Weighted SVM with Artificial Bee Colony	97.72%
[SM14]	Rough set K-means Clustering	96.49%
[SSM14]	Decision Tree	85.1%
[AYYA15]	BPDRM	97.03%
Notre travail	MLP	**98.44**%
	GA-MLP	**98.53**%
	MLP-CS	**99.02**%

TABLE 5.10 – Résultats de classification obtenues par MLP-CS et par d'autres classifieurs de la littérature sur la base WDBC.

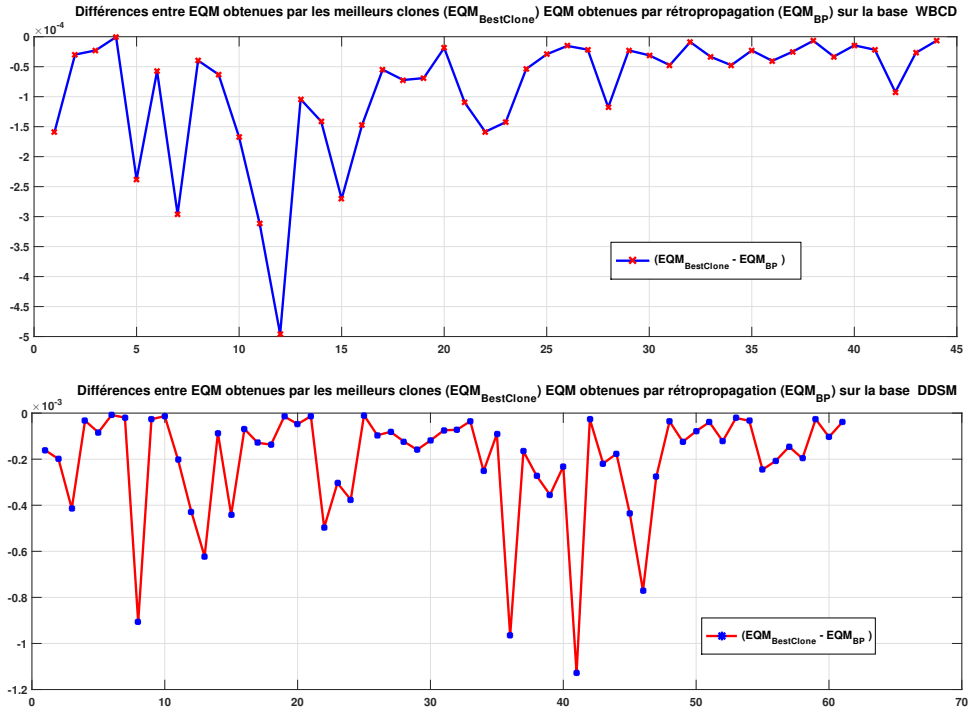


FIGURE 5.11 – Différences entre $EQM_{BestClone}$ et EQM_{BP} sur les bases de données WDBC (haut) et DDSM (bas) : toutes ces valeurs < 0 montrent que la procédure de minimisation de l'EQM est accélérée par les opérateurs du SIA.

tion, pour un diagnostic rapide et précis du cancer du sein. Nous avons présenté un PMC basé sélection clonale (MLP-CS) qui utilise les opérateurs de clonage et de mutation des SIA pour réduire les taux computationnels du PMC. Les résultats obtenus montrent que l'algorithme MLP-CS fournit les meilleurs performances. En effet, une réduction significative du temps de convergence du PMC classique a été observée, accompagnée d'une légère amélioration dans le taux de classification, et ce, sur les deux bases de données WDBC et DDSM. La comparaison avec un PMC optimisé par les algorithmes génétiques a confirmé la performance de la sélection clonale dans le domaine de l'optimisation.

Conclusion générale

Cette thèse s'inscrit dans l'objectif d'aide au diagnostic du cancer du sein par les méthodes bio-inspirées. On s'est concentré dans notre travail sur la dernière étape d'un système de Diagnostic Assisté par Ordinateur (DAO), soit la classification des masses mammaires en Bénignes / Malignes, par les Systèmes Immunitaires Artificiels (SIA). Ces algorithmes constituent un axe de recherche en cours d'exploitation qui a su s'imposer au sein de la famille des algorithmes évolutionnaires dans de nombreuses applications.

Nous avons d'abord commencé par présenter le système immunitaire naturel, tout en fournissant un aperçu sur les propriétés fondamentales qui en ont fait une modélisation informatique particulièrement avantageuses. Ensuite, nous avons tracé un état de l'art sur les principaux volets des SIA qui sont : la sélection négative, les réseaux immunitaires et la sélection clonale artificielle. Nous nous sommes focalisé dans nos travaux sur les algorithmes de sélection clonale artificielle afin d'étudier leurs performances dans le domaine du diagnostic du cancer du sein, car ces algorithmes ont l'avantage de simuler de manière explicite le processus d'une réponse immunitaire adaptative à un antigène, en utilisant deux opérateurs importants : le clonage et la mutation.

Dans l'objectif de fournir aux experts un outil rapide et performant d'aide à la décision finale ; cette thèse a visé l'apport des contributions dans deux grands volets : la reconnaissance des formes et l'optimisation.

Concernant la reconnaissance des formes, nous avons pu constater quelques remarques sur l'algorithme le plus référencé dans le domaine de la sélection clonale, à savoir l'algorithme CLONALG. Ces remarques portent sur la manière dont il est initialisé (ainsi que la plupart des algorithmes de sélection clonale), et sur l'étape de rejet des cellules mémoires et leur remplacement par d'autres aléatoirement générées afin de maintenir une bonne diversité dans l'algorithme.

Tout d'abord, et afin de traiter les problème d'initialisation, nous avons présenté une méthode simple qui permet de générer des cellules mémoires initiales représentatives de la totalité des données d'apprentissage par moyennes de sous-groupes locaux de ces dernières. Ensuite, pour éviter de rejeter des cellules potentiellement efficaces, et d'introduire l'aléatoire dans l'algorithme (qui risque de présenter de nouveaux prototypes inutiles au processus d'apprentissage), nous avons proposé trois approches dont chacune vise à maintenir une bonne diversité et à améliorer la précision en créant des cellules mémoires pertinentes.

La première approche nommée Median Filter Clonal ALGorithm (MF-CLONALG) permet de générer des cellules médianes potentielles à travers la matrice carrée des (meilleurs) cellules les plus proches de l'exemple en cours d'apprentissage, tandis-que

les deux autres approches Average Cells Clonal ALGorithm (AC-CLONALG) et Validity Interval Clonal Selection(VI-CS), consistent à créer des cellules mémoires à partir des moyennes de ces meilleurs cellules. L'approche VI-CS étant une amélioration de AC-CLONALG, diffère en l'utilisation d'un intervalle de validité pour la sélection des cellules créées afin de rejoindre l'ensemble des cellules mémoires finales utilisées dans la classification.

L'une des conclusion avec lesquelles nous nous sommes ressortis est que ces méthodes ont permis une nette amélioration des résultats de CLONALG, mais nécessitent des coûts de calcul non-négligeables. En effet, si le fait de créer et ajouter des cellules mémoires efficaces permet d'améliorer la précision et de maintenir une bonne diversité, cela exige en contre partie un temps important d'apprentissage.

A cet égard, nous avons proposé une autre approche portant sur la réduction du temps d'apprentissage, tout en préservant la précision des algorithmes de sélection clonale. L'algorithme Local Database Categorization Artificial Immune System (LDC-AIS) intègre la catégorisation par K-means et l'apprentissage des catégories par le réseau de neurones RBF au processus d'apprentissage du SIA. Cette approche utilise un mécanisme simple pour faire face aux fonctions de test coûteuses en temps, en limitant le nombre de ces tests (effectués par chaque antigène à apprendre), en outre, préserve la précision du SIA. Avec ce travail, nous avons pu atteindre l'objectif fixé, qui est le développement d'un outil rapide et efficace de reconnaissance de formes par la sélection clonale pour la classification des masses mammaires, et l'aide au diagnostic du cancer du sein.

Le deuxième volet de la thèse traite l'optimisation multimodale. En effet, les SIA, et particulièrement la sélection clonale artificielle, est largement suggérée dans la littérature pour effectuer les tâches d'optimisation. Nous nous sommes donc intéressés à explorer ces algorithmes afin d'étudier les avantages des opérateurs de clonage et de mutation dans le cadre de l'optimisation de fonctions.

Dans ce contexte, et en vue de répondre au principal inconvénient du réseau de neurones PMC (Perceptron Multi-Couches), nous avons proposé une approche d'accélération de la convergence de ce dernier tout en évitant les minima locaux. En effet, bien que le PMC soit largement référencé grâce à sa haute performance, il présente néanmoins un désavantage réputé qui est la lente convergence. Notre objectif était de trouver les poids optimaux du réseau, permettant de minimiser l'Erreur Quadratique Moyenne (EQM) et de maximiser la reconnaissance sans modifier l'algorithme de rétropropagation. Une procédure d'optimisation a été mise e œuvre, dans laquelle la rétropropagation est assistée par les processus de clonage et de mutation. L'algorithme Multi-Layer Perceptron based Clonal Selection (MLP-CS) a permis une réduction importante du temps de convergence du PMC, ainsi qu'une légère amélioration des taux de classification du cancer du sein. Par ailleurs, l'approche MLP-CS étant voisine des techniques évolutionnaires, a été comparée à un PMC basé algorithme génétique (AG). Un PMC optimisé par un AG, utilisant les mêmes paramètres, a été implémenté afin de comparer et valider les résultats obtenus par MLP-CS. Cette comparaison nous a permis de conclure que l'opérateur de croisement de l'AG nécessite plus de temps pour converger vers le seuil d'erreur comparé aux opérateurs du SIA qui sont plus rapides et plus performants. En effet, l'opérateur de clonage maintient les bons individus et le processus de mutation vise à améliorer ces derniers et augmente la probabilité de sélection des solutions les plus pertinentes, faisant du SIA un

outil plus performant que l'AG.

Bien entendu, l'exploration des classifieurs immunitaires artificiels est loin d'être achevée, il reste plusieurs points qui devraient se révéler fructueux pour des recherches futures. Par exemple, il serait intéressant de :

- S'intéresser aux deux autres volets des algorithmes immunitaires artificiels qui sont la sélection négative et les réseaux immunitaires artificiels, afin d'avoir une vue plus généralisée sur ces méthodes.
- Travailler davantage sur la réduction des coûts de calcul de ces algorithmes en appliquant des méthodes de sélection des meilleurs attributs (descripteurs).
- Expérimenter plus profondément les formules de clonage et leur impact sur la pertinence des cellules mémoires et la rapidité de l'algorithme.
- Réviser le mécanisme de mutation qui est basé sur des changements aléatoires en employant des taux de mutation relatifs à la qualité de la cellule étudiée.
- Étudier les limites du nombre de cellules mémoires générées et trouver une formule en relation avec la taille de la base d'apprentissage.
- Se pencher plus sur la capacité de généralisation des cellules mémoires et le problème de diversité de ces dernières.
- Élaborer des systèmes intelligents à travers l'hybridation des SIA avec d'autres approches d'apprentissage automatique, en profitant des avantages de chaque approche.
- Étudier davantage les performances des SIA dans des problèmes d'optimisation plus complexes.

D'une vision plus globale, notre objectif final est d'aider les cliniciens dans le diagnostic du cancer du sein. Dans cet esprit, il est possible de dire que les approches immunitaires artificielles sont adaptées à cette fin. Cependant, il existe encore des enjeux qui doivent être résolus afin que ces approches puissent être pleinement applicables, tels que :

- L'exploitation de bases de données réelles et plus volumineuses en travaillant en collaboration avec des experts.
- Association des méthodes de classification à des algorithmes robustes d'extraction se caractéristiques (étape 3 d'un système DAO).
- Pousse des limites en s'intéressant plus à la classification BIRADS (Breast Imaging-Reporting And Data System) et envisagement de la classification des microcalcification.

Annexe A

Wisconsin Diagnostic Breast Cancer (WDBC)

La base de données du Wisconsin est l'une des bases les plus référencées dans la littérature dans les applications de classification et de diagnostic du cancer du sein. Elle a été soutenue par le William H Wolberg et al. dans [WM90], et peut être téléchargée à partir de la base d'apprentissage "UCI" dans [Lic13].

WDBC se compose de données provenant de 569 cas (357 tumeurs bénignes et 212 cas de tumeurs malignes) de cytoponctions mammaires (en anglais FNA : Fine Needle Aspiration), qui consiste à prélever quelques cellules de la lésion repérée dans le sein grâce à l'aspiration par une aiguille très fine. Chaque cas contient 32 caractéristiques descriptives (descripteurs) où les deux premiers correspondent à un numéro d'identification unique de la patiente et l'état de diagnostic (B pour bénin et M pour malin).

Les 30 descripteurs restants sont calculés à partir de l'image numérisée d'une petite goutte de liquide de la masse aspirée. La matière aspirée est exprimée sur une lame de verre et colorée. L'image pour l'analyse numérique est généré par une caméra vidéo couleur montée au sommet d'un microscope et l'image est projetée dans un appareil photo. A l'aide d'une interface interactive, les modèles de contours actifs sont initialisés à proximité des limites d'un ensemble de différentes cellules. Les courbes personnalisées se déplacent et épousent la forme exacte des cellules, et une dizaine de descripteurs est calculées pour chaque cellule. Les valeurs moyenne, maximale et l'erreur standard de chaque descripteur sont calculées pour chaque image (Table A.1).

Les descripteurs sont enregistrés avec quatre chiffres significatifs, et étant donné qu'ils sont mesurés dans des échelles différentes, la fonction d'erreur sera dominée par les variables à grande échelle. Ainsi, pour éliminer l'effet des différentes échelles, la normalisation est nécessaire avant l'apprentissage [ZYL14]. Des échantillons de la base WDBC sont présentés dans la figure A.1.

Dans notre travail la base WDBC est normalisée dans l'intervalle [0, 1] selon l'équation suivante :

$$x_i = \frac{x_i^o - x_{min}}{x_{max} - x_{min}} \tag{A.1}$$

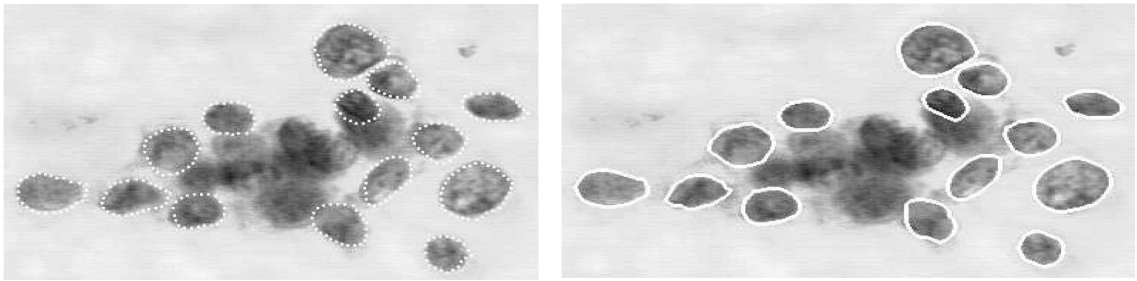


FIGURE A.1 – Images prises à l'aide d'aspiration par aiguille fine : Gauche : limites approximatives initiales des cellules, à droite : les contours après convergence vers les frontières des cellules [SWM93].

Où :
- x_i est la valeur normalisée de x_i^o ;
- x_i^o est la valeur originale du descripteur X à normaliser ;
- x_{min} la valeur minimale du descripteur X pour tout i ;
- x_{max} la valeur maximale du descripteur X pour tout i ;
- $i \in [1, NB_{WDBC}]$, avec NB_{WDBC} le nombre total des exemples de la base WDBC.

Dans les chapitres 3 et 4, le principe de validation croisée avec k=4 est utilisé (4-fold cross validation). La base WDBC est partitionnée en 4 parties égales, et chaque partie est conservée une fois pour le test quand les trois autres sont utilisée dans l'apprentissage. La moyenne des 4 résultats obtenus est considérée comme résultat d'une seule exécution. Le résultat final d'une expérimentation est la moyenne de 10 exécutions successives. Dans le chapitre 5, le même principe est utilisé mais avec une validation croisée avec k=5 (5-fold cross validation) pour des raisons comparatives.

Descripteur	Description	Moy (B - M)	Err-Std (B - M)	Max (B - M)
1.Rayon	moyenne des distances du centre au points sur le périmètre	6.98 - 28.11	0.112 - 2.873	7.93 - 36.04
2.Texture	écart-type de valeurs de gris	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
3.Périmètre	distance totale entre les points consécutifs du contour	43.79 - 188.50	0.76 - 21.98	50.41 - 251.2
4.Aire	nombre de pixels à l'intérieur du contour + 1/2 des pixels sur le périmètre	143.50 - 2501	6.80 - 542.20	185.20 - 4254
5.Régularité	différence entre la longueur d'une ligne radiale et la longueur moyenne des lignes qui l'entourent	0.053 - 0.163	0.002 - 0.135	0.027 - 1.058
6.Compacités	$\frac{périmètre^2}{aire-1}$	0.019 - 0.345	0.002 - 0.135	0.027 - 1.058
7.Concavité	gravité des portions concaves du contour : tracer des cordes entre les points non adjacents du contour et mesurer jusqu'à quel point la frontière réelle de la cellule se trouve à l'intérieur de chaque corde	0.0 - 0.427	0.0 - 0.396	0.0 - 1.252
8.Points concaves	nombre de parties concaves du contour	0.0 - 0.201	0.0 - 0.053	0.0 - 0.291
9.Symétrie	mesure de symétrie de la cellule : différence de longueur entre les lignes perpendiculaires à l'axe principal, jusqu'à la limite de la cellule	0.106 - 0.304	0.008 - 0.079	0.157 - 0.664
10.Dimension fractale	$Approximation de littoral - 1$	0.050 - 0.097	0.001 - 0.030	0.055 - 0.208

TABLE A.1 – Résumé des valeurs de descripteurs de WDBC : les deux premières colonnes représentent les noms des descripteurs et leurs descriptions. La valeur moyenne (Moy), l'erreur standard (Err-Std) et la valeur Maximale (Max) de chaque descripteur (bénins et malins) sont donnés dans trois colonnes suivantes [ZYL14].

Annexe B

Digital Database for Screening Mammography (DDSM)

La base de données numériques pour la mammographie de dépistage reconnue en anglais sous le nom (DDSM : Digital Database for Screening Mammography) a été rassemblée par un groupe de chercheurs de l'Université du sud de la Floride et a été complétée en 1991 [HB00]. Elle contient 2620 cas recueillies auprès de l'hôpital "Massachusetts General Hospital" (MGH), l'université "Wake Forest University" (WFU) et l'hôpital "Washington University of St. Louis School of Medicine" (WUSTL). DDSM a largement été utilisée par la communauté scientifique dans le domaine du diagnostic cancer du sein ; elle a l'avantage d'utiliser le même lexique normalisée par l'American College of Radiology (ACR) dans le BI-RADS (Breast Imaging-Reporting And Data System) [LSB+05] [BTN15].

Les différents dossiers des patientes ont été faits dans le cadre du dépistage du cancer du sein, et ont été classés en trois cas : cas normaux (pas de lésions), cas bénins et cas malins. Chaque fichier est composé de quatre vues contenant l'incidence oblique externe (MLO) et l'incidence cranio-caudale (CC) de chaque sein. Ces fichiers sont également fournis avec annotations données par des experts radiologues. Ces annotations permettent de décrire les différentes lésions présentes dans les images telles que le nombre et le type des anomalies (microcalcifications/masses), le résultat de la biopsie (bénin/malin), la localisation des lésions, etc.

Dans cette thèse, on traite seulement le cas de masses (pas de microcalcifications). Ces échantillons font partie de la sous-base utilisée dans nos évaluations. Une sous-base de DDSM a été créée, composée de 242 masses : 128 bénignes et 114 malignes. Ces exemples seront partitionnés (de la même manière que la base WDBC) en exemples d'apprentissage et exemples de test. La figure B.1 présente quelques échantillons de la base DDSM contenant des masses bénignes (haut) et malignes (bas).

La description des masses mammaires est une étape très importante dans un système DAO, trois nouveaux descripteurs : les points terminaux du squelette (en anglais SEP : Skeleton End Point), la sélection des protubérances (en anglais PS : Protuberance Selection) et le descripteur des masses spiculées (en anglais SMD : Spiculated Mass Descriptor) ont été proposées dans [SMMC+09], [CDM11] et [KDM12] respectivement.

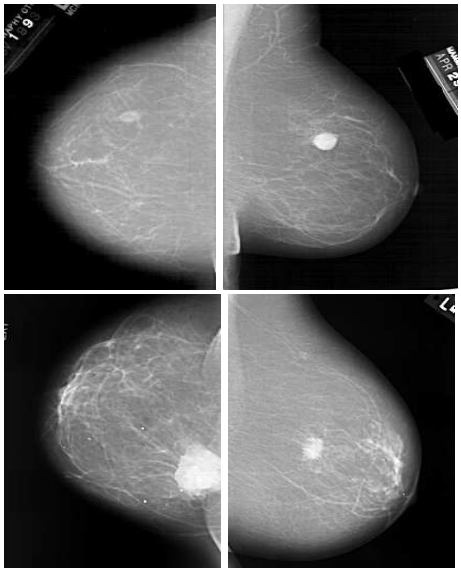


FIGURE B.1 – Échantillons de la base DDSM utilisés dans l'évaluation. Haut : cas bénins, Bas : cas malins.

Ces trois nouveaux descripteurs ont été comparés dans [Kac12] à 19 autres descripteurs proposés dans la littérature. Dans ce travail, l'ensemble des 22 attributs (B.1) est utilisé pour l'évaluation des approches proposées.

Descripteur	**Description**	**Max (B/M)**	**Mean (B/M)**	**Dev-Std (B/M)**
1.PS	la sélection des protubérances	0.175 - 1	0.078 - 0.363	0.021 - 0.230
2.SEP	les points terminaux du squelette	0.272 - 1	0.096 - 0.395	0.040 - 0.219
3.SMD	le descripteur des masses spiculées	0.284 - 1	0.126 - 0.413	0.043 - 0.174
4.A	l'aire	1 - 0.722	0.290 - 0.216	0.326 - 0.194
5.Cir	la circularité	1 - 0,560	0.747 - 0.244	0.099 - 0.137
6.Com	la compacité	0,117 - 1	0.059 - 0.264	0.021 - 0.213
7.Curv	la courbure	1 - 1	0.378 - 0.562	0.202 - 0.157
8.σ	la déviation standard de la longueur radiale normalisée	0,974 - 1	0.494 - 0.681	0.177 - 0.151
9.σ_{diff}	la différence des déviations standards	1 - 0.259	0.133 - 0.046	0.199 - 0.053
10.E	l'entropie	0.939 - 1	0.750 - 0.900	0.070 - 0.053
11.E_{diff}	l'entropie modifiée	1 - 0.867	0.500 - 0.477	0.339 - 0.199
12.D_{avg}	la moyenne de la longueur radiale normalisée	1 - 0.882	0.8335 - 0.672	0.081 - 0.085
13.NSPD	le nombre des protubérances et des dépressions importantes	0.175 - 1	0.078 - 0.363	0.021 - 0.230
14.P	le périmètre	0.471 - 1	0.181 - 0.390	0.157 - 0.276
15.A_1	le rapport de surface	0.921 - 1	0.348 - 0.573	0.157 - 0.191
16.A_2	le rapport de surface modifié	1 - 0.546	0.233 - 0.211	0.234 - 0.115
17.RECT	la rectangularité	0.974 - 1	0.496 - 0.684	0.176 - 0.149
18.M RECT	la rectangularité modifiée	1 - 0.259	0.133 - 0.046	0.199 - 0.053
19.R	la rugosité	1 - 0.525	0.314 - 0.193	0.290 - 0.115
20.ENS	le squelette elliptique normalisé	0.906 - 1	0.747 - 0.901	0.074 - 0.052
21.ZC_1	le taux de croisement en zéro	0.944 - 1	0.231 - 0.432	0.165 - 0.257
22.ZC_2	le taux de croisement en zéro modifié	1 - 0.586	0.237 - 0.2170	0.232 - 0.125

TABLE B.1 – Résumé des descripteurs de la base DDSM utilisés dans l'évaluation : les deux premières colonnes représentent les noms des descripteurs et leurs descriptions. Les valeurs Maximales , moyennes et la déviation standard de chaque attribut (bénins et malins) sont données dans trois colonnes suivantes.

Annexe C

Procédure de mise à jour des poids du MLP

La rétropropagation (backpropagation en anglais) est l'outil le plus utilisé dans les réseaux de neurones artificiels. C'est une technique qui permet de calculer le gradient de l'erreur pour chaque neurone en se basant sur l'algorithme de descente de gradient. Elle peut être appliquée à toutes les fonctions différentiables. Dans cette méthode, l'erreur de sortie du réseau est propagée dans les couches cachées, d'où le nom de rétropropagation. Le principe de rétropropagation est utilisé pour calculer $\Delta W_{j,i}$ afin de mettre à jour les poids de réseau de neurones selon l'équation :

$$W_{j,i_{New}} = W_{j,i_{Old}} + \Delta W_{j,i} \tag{C.1}$$

Dans un MLP classique avec une fonction sïgomide, la révision des poids est faite de manière récursive à partir de la dernière couche vers la première . Soit e_jl'erreur calculée pour le neurone j par :

$$e_j = D_j - Y_j \tag{C.2}$$

Où $D_j$est la sortie désirée et Y_j la sortie obtenue du neurone j calculée par :

$$Y_j = f(\sum_i W_{j,i}.X_i) \tag{C.3}$$

où X_i Correspond à l'entrée du neurone j.
Dans notre travail :
— $f(x)$ est la fonction sïgmoide définie par :

$$f(x) = \frac{1}{1+e^{-x}} \tag{C.4}$$

— L'index j représentera toujours le neurone dont le poids est à adapter,
— L'index i représentera toujours un neurone de la couche précédente par rapport au neurone j.

L'objectif est d'adapter les poids de connexions du réseau de manière à minimiser la somme des erreurs de tous les neurones de sortie. Soit E la somme des erreurs quadratiques de tous les neurones.

$$E = \sum_j (e_j)^2 = \sum_j (D_j - Y_j)^2 \tag{C.5}$$

— $f(v_j)$ est la sortie Y_j du neurone j définie par :

$$f(v_j) = f(\sum_i W_{j,i}.X_i) \tag{C.6}$$

Où $W_{j,i}$ est le poids de connexion entre le neurone i de la couche précédente, et le neurone j de la couche actuelle. Pour corriger l'erreur obtenue, on doit changer le poids $W_{j,i}$ de manière à minimiser l'erreur E. Le sens de variation de $W_{j,i}$ dépendra du gradient du cette erreur, selon l'algorithme de descente de gradient [Sny05]. Cela peut être exprimé par l'équation suivante :

$$\Delta W_{j,i} = -\alpha.\frac{\partial E}{\partial W_{j,i}} \tag{C.7}$$

Avec $0 \leq \alpha \leq 1$ représentant le pas d'apprentissage. Le nombre de poids à ajuster est le même que celui des neurones d'une même couche. En utilisant la règle de dérivation des fonctions composées définie par :

$$\frac{\partial f(x)}{\partial y} = \frac{\partial f(x)}{\partial x}.\frac{\partial x}{\partial y} \tag{C.8}$$

On peut définir :

$$\frac{\partial E}{\partial W_{j,i}} = \frac{\partial E}{\partial e_j}.\frac{\partial e_j}{\partial Y_j}.\frac{\partial Y_j}{\partial v_j}.\frac{\partial v_j}{\partial W_{j,i}} \tag{C.9}$$

• **Cas de neurone de sortie :**

$$\frac{\partial E}{\partial e_j} = \frac{\partial \sum_k (e_k)^2}{\partial e_j} = 2e_j \tag{C.10}$$

$$\frac{\partial e_j}{\partial Y_j} = \frac{\partial (D_j - Y_j)}{\partial Y_j} = -1 \tag{C.11}$$

$$\frac{\partial Y_j}{\partial v_j} = \frac{\partial (\frac{1}{1+e^{-v_j}})}{\partial v_j} = Y_j.(1 - Y_j) \tag{C.12}$$

$$\frac{\partial v_j}{\partial W_{j,i}} = \frac{(\sum_s (W_{j,s}.Y_s)}{\partial W_{j,i}} = Y_j \tag{C.13}$$

On obtient ainsi :

$$\frac{\partial E}{\partial W_{j,i}} = -2e_j.Y_i.Y_j.(1-Y_j) \tag{C.14}$$

On met : $\delta_j = e_j.Y_j.(1-Y_i)$
Et la règle de $\Delta W_{j,i}$ de la couche cachée :

$$\Delta W_{j,i} = \alpha.Y_i.\delta_j \tag{C.15}$$

• Cas d'un neurone de couche cachée :

Pour les couches cachées, il n'y a pas d'erreurs obtenues à partir de chaque neurone, et puisque le but est le même qui celui d'adapter les poids de la couche cachée, on reprend l'équation C.9 de la dérivée de l'erreur totale E par rapport à $W_{j,i}$.

— Les indices i et j désignent respectivement (comme précédemment) un neurone de la couche précédente et un neurone de la couche actuelle.
— L'index k est maintenant utilisé pour désigner un neurone de la couche suivante.

Étant donné que l'erreur e_j est inconnue, il reste :

$$\frac{\partial E}{\partial W_{j,i}} = \frac{\partial E}{\partial Y_j}.\frac{\partial Y_j}{\partial v_j}.\frac{\partial v_j}{\partial W_{j,i}} \tag{C.16}$$

Les deux derniers termes de cette équation sont les mêmes que celles de la couche de sortie (équations C.12 et C.13). Il reste à estimer le premier terme : $\frac{\partial E}{\partial Y_j}$

$$\frac{\partial E}{\partial Y_j} = 2\sum_k e_k.\frac{\partial e_k}{\partial Y_j} = 2\sum_k -e_k.[\frac{\partial Y_k}{\partial v_k}].W_{k,j} \tag{C.17}$$

En utilisant l'équation C.12 on obtient :

$$\frac{\partial E}{\partial Y_j} = -2\sum_k e_k.[Y_k.(1-Y_k)].W_{k,j}$$

On a : $\delta_k = 2e_k.Y_k.(1-Y_k)$
Alors :

$$\frac{\partial E}{\partial Y_j} = -\sum_k \delta_k.W_{k,j} \tag{C.18}$$

Par conséquent :

$$\frac{\partial E}{\partial W_{j,i}} = -Y_j.(1-Y_j).\sum_k \delta_k.W_{k,j}.Y_i \tag{C.19}$$

On met : $\delta_j = Y_j.(1-Y_j).\sum_k \delta_k.W_{k,j}$
En appliquant l'équation C.7 on obtient la règle de $\Delta W_{j,i}$ pour les couches cachées :

$$\Delta W_{j,i} = \alpha.\delta_j.Y_i \tag{C.20}$$

En utilisant les équations C.15 et C.20, on garantit une minimisation de la somme des erreurs quadratiques au niveau de chaque couche du réseau de neurones. En appliquant rétropropagation nous avons une matrice de poids ($W_{j,i_{New}}$) avec une Erreur Quadratique Moyenne inférieure à celle de l'itération précédente ($W_{j,i_{Old}}$).

Liste des publications

Papiers soumis à des journaux

Daoudi, R., & Djemal, K., *Fast and Effective Multi-Layer Perceptron Classifier Using Artificial Immune System for Breast Cancer Diagnosis*, soumis à Neural Computing and Applications (NCAA), Springer , Statut : Under Review.

Conférences internationales

1. **Daoudi, R.**, Djemal, K., & Benyettou, A., *Classification du cancer du sein par les systèmes immunitaires artificiels : Contribution à l'amélioration de l'algorithme CLONALG*, Biomedical Engineering International Conference (BIOMEIC), (2012).

2. **Daoudi, R.**, Djemal, K., & Benyettou, A., *Cells clonal selection for Breast Cancer classification*, 10^{th} International Multi-Conference on Systems, Signals & Devices (SSD), (pp. 1-4), IEEE (2013).

3. **Daoudi, R.**, Djemal, K., & Benyettou, A., *An immune-inspired approach for breast cancer classification*, 14^{th} International Conference on Engineering Applications of Neural Networks (EANN) (pp. 273-281), Springer Berlin Heidelberg (2013).

4. **Daoudi, R.**, Djemal, K., & Benyettou, A., *Digital database for screening mammography classification using improved artificial immune system approaches*, 6^{th} International Conference on Evolutionary Computation Theory and Applications (ECTA), Part of the 6^{th} International Joint Conference on Computational Intelligence (IJCCI) (pp. 244-250), (2014).

5. **Daoudi, R.**, Djemal, K., & Benyettou, A., *Using artificial immune algorithm for fast convergence of multi layer perceptron in breast cancer diagnosis application*, 10^{th} International Multi-Conference on Image Processing Theory, Tools and Applications (IPTA), (pp. 341-345), IEEE (2015).

6. **Daoudi, R.**, Djemal, K., & Benyettou, A., *Improving cells recognition by local database categorization in Artificial Immune System algorithm. Application to breast cancer diagnosis*, IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), (pp. 1-6), IEEE (2015).

7. **Daoudi, R.**, & Djemal, K., *Breast Cancer Classification by Artificial Immune Algorithm based Validity Interval Cells Selection*, accepté à : 8^{th} International Conference on Evolutionary Computation Theory and Applications (ECTA), Part of the 8^{th} International Joint Conference on Computational Intelligence (IJCCI), (Novembre 2016).

Journées de recherche

1. **Daoudi, R.**, Djemal, K., & Benyettou, A., Poster : *Digital Mammogram Database Classification using improved Artificial Immune System*, Colloque biennal « Recherche en Imagerie et Technologies pour la Santé (RITS) », (pp. 164-165), Dourdan, France, (5-27 mars 2015).
2. **Daoudi, R.**, Djemal, K., & Benyettou, A., présentation : *Classification du Cancer du Sein par les Systèmes Immunitaires Artificiels*, Journée du Gdr ISIS sur l'Analyse des images médicales pour l'aide au diagnostic (indexation, extraction de caractéristiques et reconnaissance de lésions), Université Paris-Descartes (Paris 5), France, (23 juin 2015).

Bibliographie

[AB12] Lewis J Alberts B, Johnson A. *Molecular Biology of the Cell. 4th edition. Chapter 24, The Adaptive Immune System.* Garland Science, 2012.

[ACC00] Shun-Ichi Amari, Tian Chen, and Andrzej Cichocki. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural computation*, 12(6) :1463–1484, 2000.

[AJMA07] Diego Andina, Aleksandar Jevtić, Alexis Marcano, and JM Barrón Adame. Error weighting in artificial neural networks learning interpreted as a metaplasticity model. In *Bio-inspired Modeling of Cognitive Tasks*, pages 244–252. Springer, 2007.

[AKA10] Ilhan Aydin, Mehmet Karakose, and Erhan Akin. Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection. *Expert Systems with Applications*, 37(7) :5285–5294, 2010.

[AKA11] Ilhan Aydin, Mehmet Karakose, and Erhan Akin. A multi-objective artificial immune algorithm for parameter optimization in support vector machine. *Applied Soft Computing*, 11(1) :120–129, 2011.

[AKV08] Stavros Adam, Dimitrios Alexios Karras, and Michael N Vrahatis. Revisiting the problem of weight initialization for multi-layer perceptrons trained with back propagation. In *Advances in Neuro-Information Processing*, pages 308–315. Springer, 2008.

[ATdL+02] Modupe Ayara, Jon Timmis, Rogerio de Lemos, Leandro N de Castro, and Ross Duncan. Negative selection : How to generate detectors. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, volume 1, pages 89–98. Canterbury, UK :[sn], 2002.

[AYYA15] Mohameed Sarhan Al_Duais, AbdRazak Yaakub, Nooraini Yusoff, and Faudziah Ahmed. A novel strategy for speed up training for back propagation algorithm via dynamic adaptive the weight training in artificial neural network. *Research Journal of Applied Sciences, Engineering and Technology*, 9(3) :189–200, 2015.

[B+59] Sir Frank Macfarlane Burnet et al. *The clonal selection theory of acquired immunity.* University Press Cambridge, 1959.

[Bal96] J Mark Baldwin. A new factor in evolution. *The american naturalist*, 30(354) :441–451, 1896.

[BBÖ13] Ahmet Babalik, Ismail Babaoglu, and Ahmet Özkis. A pre-processing approach based on artificial bee colony for classification by support vector machine. *International Journal of Computer and Communication Engineering*, 2(1) :68, 2013.

[Ber10] Maroun Bercachi. *Algorithme évolutionnaire à états pour l'optimisation difficile*. PhD thesis, Université Nice Sophia Antipolis, 2010.

[BH00] IA Basheer and M Hajmeer. Artificial neural networks : fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1) :3–31, 2000.

[Bro05] Jason Brownlee. Clonal selection theory & clonalg-the clonal selection classification algorithm (csca). *Swinburne University of Technology*, 2005.

[BTN15] Corinne Balleyguier and Isabelle Thomassin-Naggara. Bi-rads 2013 en mammographie : petit guide des nouveautés. *Imagerie de la Femme*, 25(1) :1–7, 2015.

[BZ93] Claudia Berek and Mike Ziegner. The maturation of the immune response. *Immunology today*, 14(8) :400–404, 1993.

[CCNS08] Angelo Ciccazzo, Piero Conca, Giuseppe Nicosia, and Giovanni Stracquadanio. An advanced clonal selection algorithm with ad-hoc network-based hypermutation operators for synthesis of topology and sizing of analog electrical circuits. In *Artificial Immune Systems*, pages 60–70. Springer, 2008.

[CCS96] Jorge Carneiro, Antonio Coutinho, and John Stewart. A model of the immune network with bt cell cooperation. ii.the simulation of ontogenesis. *Journal of Theoretical Biology*, 182(4) :531–547, 1996.

[CDM11] Imene Cheikhrouhou, Khalifa Djemal, and Hichem Maaref. Protuberance selection descriptor for breast cancer diagnosis. In *Visual Information Processing (EUVIP), 2011 3rd European Workshop on*, pages 280–285. IEEE, 2011.

[CGIR05] Felipe Campelo, Frederico G Guimarães, Hajime Igarashi, and Jaime A Ramírez. A clonal selection algorithm for optimization in electromagnetics. *Magnetics, IEEE Transactions on*, 41(5) :1736–1739, 2005.

[Che13] Li-Fei Chen. An improved negative selection approach for anomaly detection : with applications in medical diagnosis and quality inspection. *Neural Computing and Applications*, 22(5) :901–910, 2013.

[CN02] Vincenzo Cutello and Giuseppe Nicosia. Multiple learning using immune algorithms. In *Proceedings of 4th International Conference on Recent Advances in Soft Computing, RASC*, pages 102–107, 2002.

[CNNP05] Vincenzo Cutello, Giuseppe Narzisi, Giuseppe Nicosia, and Mario Pavone. An immunological algorithm for global numerical optimization. In *Artificial Evolution*, pages 284–295. Springer, 2005.

[CNP04] Vincenzo Cutello, Giuseppe Nicosia, and Mario Pavone. Exploring the capability of immune algorithms : A characterization of hypermutation operators. In *Artificial Immune Systems*, pages 263–276. Springer, 2004.

[CNP06] Vincenzo Cutello, Giuseppe Nicosia, and Mario Pavone. Real coded clonal selection algorithm for unconstrained global optimization using a hybrid inversely proportional hypermutation operator. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 950–954. ACM, 2006.

[CS92] Franco Celada and Philip E Seiden. A computer model of cellular interactions in the immune system. *Immunology today*, 13(2) :56–62, 1992.

[Das99] Dasgupta. *Artificial Immune Systems and Their Applications*. Springer Berlin Heidelberg, 1999.

[dCT02a] Leandro N de Castro and Jon Timmis. An artificial immune network for multimodal function optimization. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 699–704. IEEE, 2002.

[dCT02b] Leandro N de Castro and Jon Timmis. Hierarchy and convergance of immune networks : Basic ideas and premilinary results. In *1st International Conference on Artificial Immune Systems*, pages 231–240. Unversity of Kent at Canterbury Printing Unit, 2002.

[DCT02c] Leandro Nunes De Castro and Jonathan Timmis. *Artificial immune systems : a new computational intelligence approach*. Springer Science & Business Media, 2002.

[dCT03] Leandro Nunes de Castro and JI Timmis. Artificial immune systems as a novel soft computing paradigm. *Soft computing*, 7(8) :526–544, 2003.

[DCVZ99] Leandro Nunes De Castro and Fernando José Von Zuben. Artificial immune systems : Part i–basic theory and applications. *Universidade Estadual de Campinas, Dezembro de, Tech. Rep*, 210, 1999.

[DCVZ00] L Nunes De Castro and Fernando J Von Zuben. The clonal selection algorithm with engineering applications. In *Proceedings of GECCO*, volume 2000, pages 36–39, 2000.

[dCVZ01] L Nunes de Castro and Fernando J Von Zuben. ainet : an artificial immune network for data analysis. *Data mining : a heuristic approach*, 1 :231–259, 2001.

[DCVZ02] Leandro N De Castro and Fernando J Von Zuben. Learning and optimization using the clonal selection principle. *Evolutionary Computation, IEEE Transactions on*, 6(3) :239–251, 2002.

[Dec04] Anne Decoster. Cours de microbiologie de la faculte libre de medecine de lille / immunumogie, 2004.

[DF96] Dipankar Dasgupta and Stephanie Forrest. Novelty detection in time series data using ideas from immunology. In *Proceedings of the international conference on intelligent systems*, pages 82–87, 1996.

[Dre95] Henry Dreher. The immune power personality. *Penguin Books, Baltimore-Fabio F, Maurizio R (2006) Comparison of artificial immune systems and genetic algorithms in electrical engineering. Comput Math Electr Electron Eng*, 25(4) :792811Farmer, 1995.

[Duf09] Carine Duffaut. *Les lymphocytes du tissu adipeux humain : caractérisation et rôles*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2009.

[Enc] Encyclopédie Microsoft Encarta. Anticorps.

[Fed03] Diego Federici. Culture and the baldwin effect. In *Advances in Artificial Life*, pages 309–318. Springer, 2003.

[Flo08] Darren R Flower. Vaccines : Computational solutions. *Bioinformatics for Vaccinology*, pages 257–281, 2008.

[FLT07] Jian Fu, Zhonghua Li, and Hong-Zhou Tan. A hybrid artificial immune network with swarm learning. In *Communications, Circuits and Systems, 2007. ICCCAS 2007. International Conference on*, pages 910–914. IEEE, 2007.

[FPAC94] Stephanie Forrest, Alan S Perelson, Lawrence Allen, and Rajesh Cherukuri. Self-nonself discrimination in a computer. In *null*, page 202. Ieee, 1994.

[FPP86] J Doyne Farmer, Norman H Packard, and Alan S Perelson. The immune system, adaptation, and machine learning. *Physica D : Nonlinear Phenomena*, 22(1) :187–204, 1986.

[FT03] Alex A Freitas and Jon Timmis. Revisiting the foundations of artificial immune systems : A problem-oriented perspective. In *Artificial Immune Systems*, pages 229–241. Springer, 2003.

[FZ12] Xiaoyang Fu and Shuqing Zhang. An improved artificial immune recognition system. In *2012 IEEE International Conference on Information Science and Technology*, 2012.

[Gar05] Simon M Garrett. How do we evaluate artificial immune systems ? *Evolutionary computation*, 13(2) :145–177, 2005.

[GC04] Luis J Gonzales and James Cannady. A self-adaptive negative selection approach for anomaly detection. In *Evolutionary Computation, 2004. CEC2004. Congress on*, volume 2, pages 1561–1568. IEEE, 2004.

[GJZ10] Maoguo Gong, Licheng Jiao, and Lining Zhang. Baldwinian learning in clonal selection algorithm for optimization. *Information Sciences*, 180(8) :1218–1236, 2010.

[Glo12] Globocan. The aim of the project is to provide contemporary estimates of the incidence of, mortality and prevalence from 28 types of cancer in 184 countries worldwide., 2012.

[GN06] Hong Guo and Asoke K Nandi. Breast cancer diagnosis using genetic programming generated feature. *Pattern Recognition*, 39(5) :980–987, 2006.

[GZMJ12] Maoguo Gong, Jian Zhang, Jingjing Ma, and Licheng Jiao. An efficient negative selection algorithm with further training for anomaly detection. *Knowledge-Based Systems*, 30 :185–191, 2012.

[HB00] Michael D Heath and Kevin W Bowyer. Mass detection by relative image intensity. In *Proceedings of the 5th International Workshop on Digital Mammography (IWDM-2000)*, pages 219–225, 2000.

[HDO+98] Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4) :18–28, 1998.

[HJ08] Wenlong Huang and Licheng Jiao. Artificial immune kernel clustering network for unsupervised image segmentation. *Progress in Natural Science*, 18(4) :455–461, 2008.

[HN89] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, pages 593–605. IEEE, 1989.

[HNGS11] Norhamreeza Abdul Hamid, Nazri Mohd Nawi, Rozaida Ghazali, and Mohd Najib Mohd Salleh. Accelerating learning performance of back propagation algorithm by using adaptive gain together with adaptive momentum and adaptive learning rate on classification problems. In *Ubiquitous Computing and Multimedia Applications*, pages 559–570. Springer, 2011.

[HR02] Emma Hart and Peter Ross. Exploiting the analogy between immunology and sparse distributed memories : A system for clustering non-stationary data. In *in 1st International Conference on Artificial Immune Systems*. Citeseer, 2002.

[HS78] Ellis Horowitz and Sartaj Sahni. *Fundamentals of computer algorithms*. Computer Science Press, 1978.

[HW79] John A Hartigan and Manchek A Wong. Algorithm as 136 : A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1) :100–108, 1979.

[HYL97] P Hajela, J Yoo, and J Lee. Ga based simulation of immune networks applications in structural optimization. *Engineering Optimization*, 29(1-4) :131–149, 1997.

[IO09] Kazushi Igawa and Hirotada Ohashi. A negative selection algorithm for classification and reduction of the noise effect. *Applied Soft Computing*, 9(1) :431–438, 2009.

[JD04] Zhou Ji and Dipankar Dasgupta. Real-valued negative selection algorithm with variable-sized detectors. In *Genetic and Evolutionary Computation–GECCO 2004*, pages 287–298. Springer, 2004.

[JD09] Zhou Ji and Dipankar Dasgupta. V-detector : An efficient negative selection algorithm with probably adequate detector coverage. *Information sciences*, 179(10) :1390–1406, 2009.

[Jer74] Niels K Jerne. Towards a network theory of the immune system. In *Annales d'immunologie*, volume 125, pages 373–389, 1974.

[JTWC05] Charles A Janeway, Paul Travers, Mark Walport, and J Donald Capra. Immunobiology : the immune system in health and disease. *Immunobiology 6th edition, Taylor and Francis Group ; Garland Science*, 2005.

[Kac12] Imene Ceikhrouhou Epse Kachouri. *Description et classification des masses mammaires pour le diagnostic du cancer du sein*. PhD thesis, Ph. D. Thesis. Uiversity of Evry Val d'Essone : France, 2012.

[KDM12] Imene Cheikhrouhou Kachouri, Khalifa Djemal, and Hichem Maaref. Characterisation of mammographic masses using a new spiculated mass descriptor in computer aided diagnosis systems. *International Journal of Signal and Imaging Systems Engineering*, 5(2) :132–142, 2012.

[KDS12] Muhammad T Khan and Clarence W De Silva. Autonomous and robust multi-robot cooperation using an artificial immune system. *International Journal of Robotics and Automation*, 27(1) :60, 2012.

[KPST08] Nitesh Khilwani, Anoop Prakash, Ravi Shankar, and MK Tiwari. Fast clonal algorithm. *Engineering Applications of Artificial Intelligence*, 21(1) :106–128, 2008.

[Kra14] Oliver Kramer. *A Brief Introduction to Continuous Evolutionary Optimization*. Springer, 2014.

[KSJ+00] Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[KT03] Johnny Kelsey and Jon Timmis. Immune inspired somatic contiguous hypermutation for function optimisation. In *Genetic and Evolutionary Computation, GECCO 2003*, pages 207–218. Springer, 2003.

[Let07] Philippe Letonturier. *Immunologie générale*. 8e édition, Elsevier Masson, 2007.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[LLS08] Lu Liu, Wanyu Liu, and Xiaoming Sun. Automated detection of pulmonary nodules in ct images with support vector machines. In *International Symposium on Instrumentation Science and Technology*, pages 713326–713326. International Society for Optics and Photonics, 2008.

[LSB+05] L Lévy, M Suissa, J Bokobsa, H Tristant, J-F Chiche, B Martin, and G Teman. Présentation de la traduction française du bi-rads®(breast imaging reporting system and data system). *Gynécologie obstétrique & fertilité*, 33(5) :338–347, 2005.

[Lv07] Jia Lv. Study on chaos immune network algorithm for multimodal function optimization. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 3, pages 684–689. IEEE, 2007.

[Mac67] J MacQeen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967*, volume 1, pages 281–297. University of California Press, 1967.

[MAC+92] Mohamad T Musavi, Wahid Ahmed, Khue Hiang Chan, Kathleen B Faris, and Donald M Hummels. On the training of radial basis function classifiers. *Neural networks*, 5(4) :595–603, 1992.

[mdlsO13] Organisation mondiale de la santé (OMS). Derniéres statistiques mondiales sur le cancer, en augmentation á 14,1 millions de nouveaux cas en 2012 :

L'augmentation marquée du cancer du sein demande des réponses, 12 décembre 2013.

[MH09] Chris McEwan and Emma Hart. Representation in the (artificial) immune system. *Journal of Mathematical Modelling and Algorithms*, 8(2) :125–149, 2009.

[MKS+14] Prasant Kumar Mahapatra, Mandeep Kaur, Spardha Sethi, Rishabh Thareja, Amod Kumar, and Swapna Devi. Improved thresholding based on negative selection algorithm (nsa). *Evolutionary Intelligence*, 6(3) :157–170, 2014.

[MP43] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4) :115–133, 1943.

[MV] Virginie Marquet and Pierre Viora. Cours terminale s svt.

[N+02] Mark Neal et al. An artificial immune system for continuous analysis of time-varying data. In *1st International Conference on Artificial Immune Systems (ICARIS), University of Kent at Canterbury, UK*. DTIC Document, 2002.

[Nag14] ZA Nagy. *History of modern immunology : the path toward understanding*. Amsterdam [etc.] : Elsevier, 2014.

[NUCG03] Olfa Nasraoui, Cesar Cardona Uribe, Carlos Rojas Coronel, and Fabio Gonzalez. Tecno-streams : tracking evolving clusters in noisy data streams with a scalable immune system learning model. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 235–242. IEEE, 2003.

[Pai81] Emilie Pailloux. *Organisation du système immunitaire félin*. PhD thesis, UNIVERSITE CLAUDE-BERNARD-LYON I, 1981.

[Pen14] Lucile Penaud. *Principales innovations biotechnologiques dans la production des vaccins anti-tumoraux*. PhD thesis, Universitè de Poitiers, France, 2014.

[Per89] Alan S Perelson. Immune network theory. *Immunological reviews*, 110(1) :5–36, 1989.

[PISM13] Ayi Purbasari, Supriana S Iping, O Setiono Santoso, and Rila Mandala. Designing artificial immune system based on clonal selection : Using agent-based modeling approach. In *Modelling Symposium (AMS), 2013 7th Asia*, pages 11–15. IEEE, 2013.

[PL15] Yong Peng and Bao-Liang Lu. Hybrid learning clonal selection algorithm. *Information Sciences*, 296 :128–146, 2015.

[Pow87] Michael JD Powell. Radial basis functions for multivariable interpolation : a review. In *Algorithms for approximation*, pages 143–167. Clarendon Press, 1987.

[RB93] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning : The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. IEEE, 1993.

[RHL03] Liu Ruochen, Du Haifeng, and Jiao Licheng. Immunity clonal strategies. In *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, pages 290–295. IEEE, 2003.

[RHW85] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[SA13] Catherine Gervais St-Amour. Étude de la différenciation des lymphocytes b mémoires en milieu sans sérum. Master's thesis, Universitè LAVAL, Quèbec, Canada, 2013.

[Sak84] Michel Sakarovitch. *Optimisation combinatoire : Programmation discrète*, volume 2. Editions Hermann, 1984.

[SD16] Guilherme Costa Silva and Dipankar Dasgupta. A survey of recent works in artificial immune systems. In *HANDBOOK ON COMPUTATIONAL INTELLIGENCE : Volume 2 : Evolutionary Computation, Hybrid Systems, and Applications*, pages 547–586. World Scientific, 2016.

[SFT03] Andrew Secker, Alex Alves Freitas, and Jon Timmis. Aisec : an artificial immune system for e-mail classification. In *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 1, pages 131–138. IEEE, 2003.

[SGHL06] Zhuo-Yue Song, XZ Gao, Xian-Lin Hiang, and HS Lin. A modified immune optimization algorithm. In *Machine Learning and Cybernetics, 2006 International Conference on*, pages 2184–2189. IEEE, 2006.

[SJW92] W Schiffmann, M Joost, and R Werner. Optimization of the backpropagation algorithm for training multilayer perceptrons. *Univ. Koblenz, Inst. Physics, Rheinau*, pages 3–4, 1992.

[SM14] T Sridevi and A Murugan. An intelligent classifier for breast cancer diagnosis based on K-Means clustering and rough set. *International Journal of Computer Applications*, 85(11) :38–42, 2014.

[SMMC+09] Dorra Sellami-Masmoudi, Hichem Maaref, Imen Cheikhrouhou, Khalifa Djemal, and Nabil Derbel. Empirical descriptors evaluation for mass malignity recognition. In *First International Workshop on Medical Image Analysis and Description for Diagnosis Systems (MIAD 2009)*, pages 91–100, 2009.

[Sny05] Jan Snyman. *Practical mathematical optimization : an introduction to basic optimization theory and classical and new gradient-based algorithms*, volume 97. Springer Science & Business Media, 2005.

[SS11] Anurag Sharma and Dharmendra Sharma. Clonal selection algorithm for classification. In *Artificial Immune Systems : 10th International Conference, ICARIS 2011, Cambridge, UK, July 18-21, 2011. Proceedings*, volume 6825, page 361. Springer, 2011.

[SSM14] M Sadhana, A Sankareswari, and M MCA. A proportional learning of classifiers using breast cancer datasets. *International Journal of Computer Science and Mobile Computing*, 3(11) :223–232, 2014.

[Ste98] Jeanne Mager Stellman. *Encyclopaedia of occupational health and safety*. International Labour Organization, 1998.

[Ste06] Tamara Stern. Set cover problem. Technical report, MIT Matimatics, 2006.

[SWM93] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE's Symposium on Electronic Imaging : Science and Technology*, pages 861–870. International Society for Optics and Photonics, 1993.

[Tab08] Georges Tabarani. *DC-SIGN, un récepteur détourné par des nombreux pathogènes : carcaterisation biochimique, structurale et développement d'inhibiteurs*. PhD thesis, Université de Grenoble, 2008.

[Tim13] Jon Timmis, 2013.

[TN01] Jon Timmis and Mark Neal. A resource limited artificial immune system for data analysis. *Knowledge-Based Systems*, 14(3) :121–130, 2001.

[TNH00] Jon Timmis, Mark Neal, and John Hunt. An artificial immune system for data analysis. *Biosystems*, 55(1) :143–150, 2000.

[Ton83] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909) :575–581, 1983.

[TRR12] Mashrura Tasnim, Shahriar Rouf, and Md Saifur Rahman. A clonalg-based approach for the set covering problem. In *Computer and Information Technology (ICCIT), 2012 15th International Conference on*, pages 42–49. IEEE, 2012.

[TRR14] Masruba Tasnim, Shahriar Rouf, and M Sohel Rahman. A clonalg-based approach for the set covering problem. *Journal of Computers*, 9(8) :1787–1795, 2014.

[Tur97] Peter Turney. How to shift bias : Lessons from the baldwin effect. *Evolutionary Computation*, 4(3), 1997.

[VCDV88] Francisco Varela, Antonio Coutinho, Bruno Dupire, and Nelson Vaz. Cognitive networks : immune, neural and otherwise. *Theoretical immunology*, 2 :359–375, 1988.

[VdCMVZ03] Patrícia A Vargas, Leandro N de Castro, Roberto Michelan, and Fernando J Von Zuben. An immune learning classifier network for autonomous navigation. In *Artificial Immune Systems*, pages 69–80. Springer, 2003.

[VG10] Bram Van Ginneken. Computer-aided diagnosis in chest imaging : how to improve performance and avoid reinventing the wheel. In *Proceedings of the 2010 IEEE international conference on Biomedical imaging : from nano to Macro*, pages 274–274. IEEE Press, 2010.

[VMR+88] Thomas P Vogl, JK Mangis, AK Rigler, WT Zink, and DL Alkon. Accelerating the convergence of the back-propagation method. *Biological cybernetics*, 59(4-5) :257–263, 1988.

[VSKT10] Harsh Vazirani, Anupam Shukla, R Kala, and R Tiwari. Diagnosis of breast cancer by modular neural network. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, volume 7, pages 115–119. IEEE, 2010.

[Wat01] Andrew B Watkins. Airs : A resource limited artificial immune classifier. Master's thesis, Mississippi State University, 2001.

[Wat05] Andrew B Watkins. *Exploiting immunological metaphors in the development of serial, parallel, and distributed learning algorithms*. PhD thesis, University of Kent at Canterbury, 2005.

[WCA14] Kung-Jeng Wang, Kun-Huang Chen, and Melani-Adrian Angelia. An improved artificial immune recognition system with the opposite sign test for feature selection. *Knowledge-Based Systems*, 71 :126–145, 2014.

[Wei09] Thomas Weise. Global optimization algorithms-theory and application. *Self-Published,*, 2009.

[WG03] Jennifer White and Simon Garrett. Improved pattern recognition with artificial clonal selection ? *Artificial Immune Systems*, pages 181–193, 2003.

[WM90] William H Wolberg and Olvi L Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences*, 87(23) :9193–9196, 1990.

[WT04a] Andrew Watkins and Jon Timmis. Artificial immune recognition system (airs) : Revisions and refinements. In *AISB 2004 Convention*, page 18, 2004.

[WT04b] Andrew Watkins and Jon Timmis. Exploiting parallelism inherent in airs, an artificial immune classifier. In *Artificial immune systems*, pages 427–438. Springer, 2004.

[WTB04] Andrew Watkins, Jon Timmis, and Lois Boggess. Artificial immune recognition system (airs) : An immune-inspired supervised learning algorithm. *Genetic Programming and Evolvable Machines*, 5(3) :291–317, 2004.

[XLT05] Ji-Qing Xian, Feng-Hua Lang, and Xian-Lun Tang. A novel intrusion detection method based on clonal selection clustering algorithm. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3905–3910. IEEE, 2005.

[YH04] Ying Yu and Chao-Zhen Hou. A clonal selection algorithm by using learning operator. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 5, pages 2924–2929. IEEE, 2004.

[Zha11] Yichi Zhang. *Distributed Intrusion Detection System in A Multi-Layer Network Architecture of Smart Grids*. PhD thesis, The University of Toledo, 2011.

[ZQT13] Jinquan Zeng, Zhiguang Qin, and Weiwen Tang. Anomaly detection using a novel negative selection algorithm. *Journal of Computational and Theoretical Nanoscience*, 10(12) :2831–2835, 2013.

[ZYL14] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4) :1476–1482, 2014.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Classification du cancer du sein par des approches basées sur les systèmes immunitaires artificiels.

Mots clés : Cancer du sein, Classification, Systèmes Immunitaires Artificiels, Sélection, Clonage, Mutation.

Le cancer du sein arrive dans le monde en première position en termes d'incidence et de mortalité parmi les différentes localisations cancéreuses chez les femmes. Malgré les avancées significatives faites ces dernières décennies en vue d'améliorer la gestion de ce type de cancer, des outils de diagnostic plus précis sont encore nécessaires pour aider les experts à lutter contre cette maladie mortelle. Dans ce cadre, des travaux de recherche considérables ont été réalisés dans l'espoir d'apporter de nouvelles perspectives pour l'amélioration du diagnostic du cancer du sein, en développant des systèmes de Décision Assistés par Ordinateur (DAO). Beaucoup de travaux se sont dirigés vers la détection de la présence de tissus cancéreux dans le sein et la classification de tumeurs en utilisant des outils issus de l'intelligence artificielle souvent inspirés par les systèmes naturels. En l'occurrence, les Systèmes Immunitaires Artificiels (SIA) constituent un domaine de recherche qui comble les domaines de l'immunologie, l'informatique et l'ingénierie. Les principaux développements au sein des systèmes immunitaires artificiels, ont mis l'accent sur trois principales théories immunologiques : la sélection clonale, les réseaux immunitaires et la sélection négative. Nous nous intéressons dans ce travail à l'utilisation de algorithmes de sélection clonale pour la classification des cellules mammaires en Bénignes/Malignes. En effet, ces approches sont généralement basées sur deux principaux processus : la reconnaissance de la forme de l'antigène et la sélection de la cellule mémoire spécifique à ce dernier. L'idée établie est que seules les cellules mémoires capables de reconnaître l'antigène sont sélectionnées pour le clonage et la mutation. Après avoir présenté le principe de ces algorithmes, nous étudierons, à travers plusieurs approches, leurs performances. Tout d'abord, on s'intéresse à l'amélioration de l'algorithme CLONALG, qui est un des algorithmes de base dans le domaine de la sélection clonale artificielle.

Afin de renforcer l'apprentissage de ce dernier, avec une meilleure initialisation et une diversité maîtrisée, trois différentes méthodes sont proposées appelées Median Filter Clonal ALGorithm (MF-CLONALG), Average Cells Clonal ALGorithm (AC-CLONALG) et Validity Interval Clonal Selection (VI-CS). Cependant, bien qu'elles soient performantes, ces approches nécessitent un temps important de calcul. Dans ce contexte, la seconde approche qu'on propose vise à réduire les taux de calcul de ces algorithmes (et ceux des SIA en général) sans affecter leurs performances. L'algorithme Local Database Categorization Artificial Immune System (LDC-AIS) utilise le regroupement par K-means pour la catégorisation locale des données, et le réseau de neurones RBF pour l'apprentissage des catégories, afin d'accélérer le processus de sélection. La dernière partie de la thèse est dédiée à l'optimisation multimodale. En effet, après avoir présenté les algorithmes de sélection clonale comme outil compétitif de reconnaissance des formes et de classification, nous nous sommes intéressés à explorer ce concept, afin de démontrer les avantages des opérateurs de clonage et de mutations dans le cadre de l'optimisation de fonctions. En réponse à certains inconvénients du réseau de neurones MLP (Multi-Layer Perceptron), une procédure d'optimisation en plusieurs étapes est proposée, dans laquelle la rétropropagation est assistée par les processus de clonage et de mutation, pour une convergence plus rapide et précise du MLP. L'approche Multi-Layer Perceptron based Clonal Selection (MLP-CS) étant voisine des techniques évolutionnaires est comparée à un MLP optimisé par un algorithme génétique. Chacune des approches proposées dans ce travail est testée et comparée à différents travaux antérieurs en utilisant deux différentes bases de données mammaires à savoir la Wisconsin Diagnostic Breast Cancer (WDBC) et la Digital Database for Screening Mammography (DDSM).

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Breast cancer classification using artificial immune system approaches.

Keywords : Breast Cancer, Classification, Artificial Immune Systems, Selection, Cloning, Mutation.

Breast cancer arrives in the world in first place in terms of incidence and mortality among the different cancer localizations in women. Despite the significant progress made in recent decades to improve the management of this type of cancer, more accurate diagnostic tools are still necessary to help experts fight against this fatal disease. In this context, considerable research studies have been carried out to bring new perspectives for the improvement of the diagnosis of breast cancer, by developing Computer-Aided Diagnosis systems (CAD). Many works were directed to detecting the presence of cancerous tissues in the breast and tumor classification using tools from artificial intelligence often inspired by natural systems. In this case, Artificial Immune Systems (AIS) are a research field that bridges the fields of immunology, computer science and engineering. The main developments in artificial immune systems, have focused on three main immunological theories: clonal selection, immune networks and negative selection. We focus in this work on the use of clonal selection algorithms for classification of breast cells in Benign / Malignant. Indeed, these approaches are generally based on two main processes: the shape recognition of the antigen and selection of the specific memory cell to it. The established idea is that only memory cells capable of recognizing the antigen are selected for cloning and mutation. After introducing the principle of these algorithms we will study, through various approaches, their performances. First, we focus on improving CLONALG algorithm, which is a basic algorithm in the field of artificial clonal selection. To enhance the learning of the latter, with a better initialization and controlled diversity, three different methods are proposed appointed Median Filter Clonal ALGorithm (MF-CLONALG), Average Cells Clonal ALGorithm (AC-CLONALG) and Validity Interval Clonal Selection (VI-CS).

However, although successful, these approaches require significant computing time. In this context, the second proposed approach aims at reducing the computational rates of these algorithms (and those of the AIS in general) without affecting their performance. The Local Database Categorization Artificial Immune System algorithm (LDC-AIS) uses clustering by K-means for local data categorization, and RBF neural network for learning categories, to accelerate the selection process. The last part of the thesis is dedicated to multimodal optimization. Indeed, after having presented the clonal selection algorithms as competitive tools of pattern recognition and classification, we were interested in exploring this concept, to demonstrate the benefits of cloning and mutation operators in functions optimization's framework. In response to some drawbacks of the MLP neural network (Multi-Layer Perceptron), an optimization procedure in several stages is proposed, in which the back-propagation is assisted by cloning and mutation processes, for fast and accurate convergence of MLP. Being close to evolutionary techniques, the Multi-Layer Perceptron based Clonal Selection approach (MLP-CS) is compared to an MLP optimized by a genetic algorithm. Each of the approaches proposed in this work is tested and compared to different previous works using two different breast databases which are the Wisconsin Diagnostic Breast Cancer (WDBC), and the Digital Database for Screening Mammography (DDSM).

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France