IV.1.2 Description de la base de la maladie du cancer du sein

La base de données du cancer du sein dénommée « Wisconsin Breast Cancer Database » a été obtenue par l'Université du Wisconsin [50]

elle contient les informations médicales de 699 cas cliniques relatifs au cancer du sein classés comme bénin ou malin : 458 patientes (soit 65.5%) sont des cas bénins et 241 patientes (soit 34.5%) sont des cas malins.

La base de données contient 16 données manquantes; les patientes sont caractérisées par 11 attributs : le premier fait référence à l'identificateur de la patiente et le dernier représente la classe: le diagnostic est de 2 si le cas est bénin ,4 si le cas est malin quant aux 9 autres, ils représentent des cas cliniques suivants:

1-Clump Thickness: l'épaisseur de la membrane plasmique d'une cellule cancéreuse est plus importante que celle d'une cellule normale.

2. Uniformity of Cell Size : les cellules cancéreuses sont caractérisées par une anisocytose, à savoir une inégalité au niveau de la taille par comparaison avec les cellules saines.

3. Uniformity of Cell Shape : les cellules cancéreuses sont marquées par des contours irréguliers ainsi que des incisures

4. Shape Marginal Adhesion: une surexpression de la protéine integrin beta3 au niveau de la surface de la cellule cancéreuse.

5 .Single Epithelial Cell Size: étant donné que les cellules épithéliales sont absentes à l'état naturel au niveau de la moelle osseuse et qu'elles ne sont pas détectées chez les individus sains, la moelle osseuse peut, de ce fait, être considérée comme un indicateur de maladie métastatique chez les patients atteints du cancer du sein au stade primaire.

6. Bare Nuclei: à l'état normal, les nucléoles se trouvent à l'intérieur du noyau. Dans le cas où ses derniers se trouvent confondus avec le cytoplasme cela indique que la cellule présente une anomalie et qu'elle est susceptible de devenir cancéreuse.

7 Bland Chromatin : H2az est une protéine qui induit l'expression du gène du récepteur d'œstrogènes.

La surproduction de cette protéine est un marqueur de présence de cellules cancéreuses au niveau du sein étant donné qu'elles son hormono-dépendantes.

8 Normal Nucleoli : L'ADN est naturellement protégé par une membrane nucléraire. Une défaillance observée au niveau de cette membrane peut refléter une croissance tumorale.

9. Mitoses : La mitose est un processus de division cellulaire régulé permettant de reproduire des cellules filles génétiquement identique à la cellule parentale.

Les cellules malignes sont caractérisées par une division cellulaire anarchique et intense par comparaison avec une population cellulaire normale.

Remarque : Etant donné qu'il y a 16 données manquantes, nous nous sommes restreints à travailler sur 683 /699 patientes

IV.1.3 Les outils de programmation

Afin de réaliser ce travail nous avons utilisé des outils suivants :

1) pour effectuer la phase d'apprentissage et la phase de test sur la base de données nous avons utilisé le langage Matlab 7

2) pour illustrer le principe des systèmes multi agents nous avons sollicité la plateforme jade en ayant recours au langage Java

**A noter que le passage de Matlab à Java s'est réalisé à l'aide de la bibliothèque jMatlink*

IV.1.4 principe de la classification

Dans certains cas, il est possible de décrire complètement, de manière linguistique, la démarche de classification; dans ce cas, un algorithme reproduisant cette démarche peut être construit et le problème est résolu. Dans d'autres cas, il est impossible de décrire précisément la classification; une solution consiste alors à demander à un professeur (expert) de classer un échantillon d'objets.

Des méthodes de résolution, qui apprennent par l'exemple, sont capables de reproduire la classification de l'expert et, ensuite, de classer automatiquement de nouveaux exemples inconnus.

Nous avons utilisé ce deuxième type qui se base sur l'apprentissage par l'exemple.
Pour effectuer la phase d'apprentissage il est nécessaire d'avoir une optimisation de la fonction de coût d'où la définition de cette dernière est primordiale.
Car celle-ci sert à mesurer l'écart entre la sortie du modèle et les mesures faites sur les exemples d'apprentissage. Ainsi elle consiste à avoir une maximisation du taux de reconnaissance et en même temps une minimisation de l'erreur.

L'idée est de disposer d'un ensemble permettant de tester la qualité de la procédure de classification induite. On partitionne l'échantillon en un ensemble d'apprentissage et un ensemble test. La répartition entre les deux ensembles doit être faite expérimentalement. L'estimation de l'erreur réelle est alors l'erreur apparente mesurée sur l'ensemble test.
La qualité de l'apprentissage augmente avec la taille de l'ensemble d'apprentissage. Mais, dans la pratique, la taille de l'échantillon est limitée.
Cette méthode donne de bons résultats lorsque l'échantillon est "assez" grand. Il existe peu de résultats théoriques sur les tailles d'échantillon nécessaires pour utiliser cette méthode, nous ne disposons que de résultats empiriques qui dépendent du problème (souvent, plusieurs centaines d'exemples).
La répartition de l'échantillon entre les deux ensembles se fait en général dans des proportions 1/2, 1/2 pour chacun des deux ensembles ou 2/3 pour l'ensemble d'apprentissage et 1/3 pour l'ensemble test.

IV.1.5 Phase de test (d'évaluation)

Cette phase doit permettre l'affectation d'un nouvel objet à l'une des classes, au moyen d'une règle de décision intégrant les résultats de la phase d'apprentissage. L'objectif est d'obtenir une estimation la plus fidèle possible du comportement du classifieur dans des conditions réelles d'utilisation. Pour cela, des critères classiques comme les taux de classification et les taux d'erreur sont presque systématiquement utilisés. Mais d'autres critères, comme la spécificité et la sensibilité, apportent aussi des informations utiles.

a-**taux de classification** : Les taux de classification et d'erreurs permettent d'évaluer la qualité du classifieur par rapport au problème pour lequel il a été conçu. Ces taux sont évalués grâce à une base de test qui contient des formes étiquetées par leur classe réelle d'appartenance comme celles utilisées pour l'apprentissage afin de pouvoir vérifier les réponses du classifieur.

Pour que l'estimation du taux de reconnaissance soit la plus fiable possible, il est important que le classifieur n'ait jamais utilisé les échantillons de cette base pour faire son apprentissage, de plus cette base de test doit être suffisamment représentative du problème de classification.

En général, quand les échantillons étiquetés à disposition sont suffisamment nombreux, ils sont séparés en deux parties disjointes et en respectant les proportions par classes de la base initiale. Une partie sert pour former la base d'apprentissage et l'autre pour former la base de test.

Les performances en termes de taux de classification sont alors déterminées en présentant au classifieur chacun des exemples de la base de test et en comparant la classe donnée en résultat à la vraie classe.

Le taux de classification correcte est défini par :

$$CC = VP(i)+VN(i) / VP(i)+VN(i)+FP(i)+FN(i)$$

Avec : *VP* : Vrai Positif : nombre de positifs classés positifs.
VN : Vrai Négatif : nombre de négatifs classés négatifs.
FP : Faux Positif : nombre de négatifs classés positifs.
FN : Faux Négatif : nombre de positifs classés négatifs.

b. sensibilité et spécificité

Sensibilité : on appelle sensibilité (Se) du test sa capacité de donner un résultat positif quand la maladie est présente. Dans le langage des probabilités, la sensibilité mesure la probabilité conventionnelle que le test soit positif lorsque la maladie est présente. La sensibilité est estimée par la proportion de résultats positifs par suite de l'application du test à un groupe d'individus reconnus comme ayant la maladie.

$$Se = VP/VP + FN$$

Spécificité : on appelle spécificité (Sp) du test cette capacité de donner un résultat négatif quand la maladie est absente. Dans le langage des probabilités, la spécificité mesure la probabilité conventionnelle que le test soit négatif lorsque la maladie est absente. La spécificité est estimée par la proportion de résultats négatifs conséquemment à l'application du test à un groupe d'individus reconnus comme n'ayant pas la maladie.

$$Sp = VN/VN + FP$$

IV.2 TRAVAIL EFFECTUE

Ce travail se divise en deux parties,

- la première partie étant la classification par les trois méthodes (méthode neuronale, neuro-génétique paramétrique, neuro-génétique structurel)
- la seconde décrit comment nous avons procédé pour illustrer le principe multi agents et ceci à l'aide des outils cités précédemment.

IV.2.1 Classification mono-agents

Dans ce qui suit nous détaillons les démarches suivies afin de réaliser la classification

IV.2.1.1 Classification neuronale (CNC)

- ***Implémentation d'un RNMC (réseau de neurone multicouches)***

✓Apprentissage structurel

✓Nombre de neurones d'entrée : nous avons utilisé 9 neurones (les 9 vecteurs d'entrées), nous avons jugé d'enlever le premier(le code de la patiente) vu qu'il n'a pas un impact sur les résultats

✓Nombre de neurones cachés : après plusieurs expérimentations, nous avons fixé le nombre de neurones à 15

✓Nombre de neurones de sortie : nous avons utilisé un seul neurone représentant la classe

Remarque : La fonction d'activation pour les neurones cachés est 'logsig'

La fonction d'activation pour la couche de sortie est 'purelin'

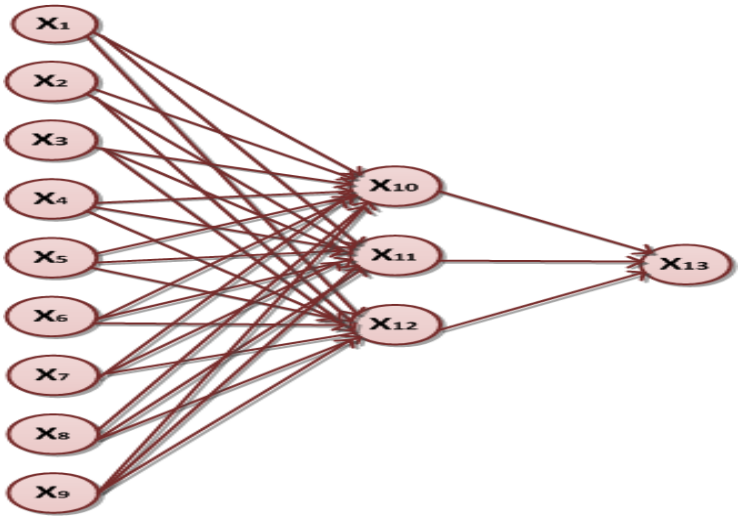


Figure IV.22 : Architecture du réseau RNMC

✓**Apprentissage paramétrique**

•***Choix de l'algorithme d'apprentissage***

L'algorithme d'apprentissage utilisé c'est l'algorithme de rétro propagation par la méthode de Levenberg-Marquardt (LM)

Cet algorithme est utilisé dans les réseaux de type feed forward, ce sont des réseaux de neurones à couches, ayant une couche d'entrée, une couche de sortie, et au moins une couche cachée. Il n'y a pas de récursivité dans les connexions, et pas de connexions entre neurones de la même couche. Le principe de la rétro propagation consiste à présenter au réseau un vecteur d'entrées, de procéder au calcul de la sortie par propagation à travers les couches, de la couche d'entrée vers la couche de sortie en passant par les couches cachées. Cette sortie obtenue est comparée à la sortie désirée, une erreur est alors obtenue. A partir de cette erreur, est calculé la sortie qui est à son tour propagé de la couche de sortie vers la couche d'entrée, d'où le terme de rétro propagation. Cela permet la modification des poids du réseau et donc l'apprentissage. L'opération est réitérée pour chaque vecteur d'entrée et cela jusqu'à ce que le critère d'arrêt soit vérifié.

IV.2.1.2 Classifieur neuro-génétique

Pour ce classifieur, nous avons débuté par la même architecture que le RNMC

✓**Apprentissage paramétrique**

En premier lieu une population initiale est crée (taille de la population fixé expérimentalement, en ce qui nous concerne nous avons utilisé N=100) avec des poids aléatoires compris entre -1,0 à + 1,0.

une fois le jeu de poids généré il sera représenté sous la forme d'un chromosome

Sachant que le chromosome est une collection de gènes (voir chapitre 3); et qu'en ce qui nous concerne nous avons gardé la structure suivante (9-3-1) avec 30 connections (liens pondérés entre un neurone et un autre), le chromosome sera alors représenté par 30 gènes.

Maintenant, nous évaluons les performances de chaque chromosome en définissant la fonction fitness ; dans notre cas nous avons utilisé le calcul de l'erreur quadratique

Selon les résultats obtenus, les meilleurs chromosomes (ayant l'erreur quadratique la plus petite) sont sélectionnés (Stochastique Uniform), mutés(Gaussian) et croisés (Pc=0.8) formant ainsi une nouvelle population
La procédure se répète jusqu'à atteindre le nombre de génération (N=100)

1	2	3	4	5	6	7	8	9	10
-0.6041	-0.7186	-0.4667	0.3974	-0.6138	-0.8826	1.5129	0.1252	3.2231	-0.4317
11	12	13	14	15	16	17	18	19	20
2.5582	1.2547	4.1567	-0.6688	2.1089	2.1227	1.4714	2.1566	-0.5217	0.4696
21	22	23	24	25	26	27	28	29	30
1.5524	-0.7317	-0.5079	0.7260	-0.5946	0.9116	-2.3084	-0.3190	2.0591	0.5714

Figure IV.23 : chromosome des poids synaptiques

✓***Apprentissage structurel*** il suit le même principe que pour l'apprentissage paramétrique, sauf que nous avons :

1) remplacer le poids par une connexion binaire : 1 s'il y a une connexion entre un neurone et un autre sinon 0.
b) dans ce cas nous avons gardé la structure (9-4-1) avec 40 connexions, (Figure IV.24)

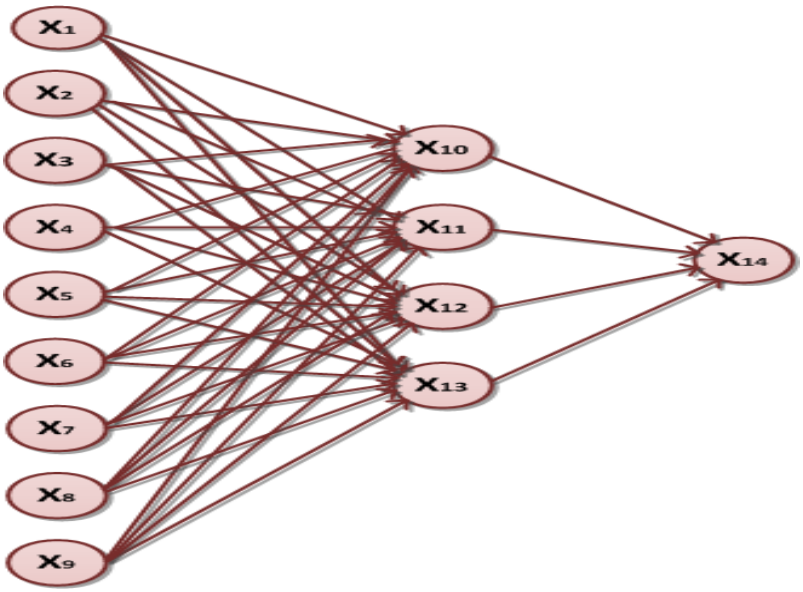


Figure IV.24 architecture du réseau neuro génétique avant apprentissage

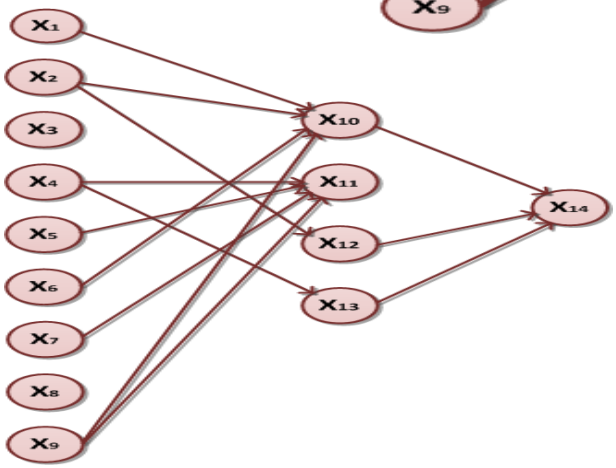


Figure IV.25: Architecture du réseau neuro génétique après apprentissage

1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	1	0	0	0	0
11	12	13	14	15	16	17	18	19	20
0	1	1	0	1	0	1	0	0	1
21	22	23	24	25	26	27	28	29	30
0	0	0	0	0	0	0	0	0	0
31	32	33	34	35	36	37	38	39	40
1	0	0	0	0	0	1	0	1	1

Figure IV.26: Chromosome des connexions

IV.2.1.3 Résultats

Pour effectuer la répartition des 683 patientes : nous avons gardé 2/3(456) pour la phase d'apprentissage et 1/3 (227) pour la phase de test

➢ **Résultats obtenus par le classifieur neuronal**

Erreur atteinte du réseau	Nombre de neurones cachés	nombre d'itération	CC	Se	Sp	VP	VN	FP	FN
0.01	15	150	98,65%	96%	99,42%	48	172	1	2

Tableau IV. 1 : Résultats du classifieur neuronal (CRN)

➢ **Résultats obtenus par le classifieur neuro-génétique**

✓ **Apprentissage paramétrique**

Population	génération	Nombres de neurones cachés	CC	Se	Sp	VP	VN	FP	FN
100	100	3	97,23%	97,07%	98,28%	51	172	3	1

Tableau IV.2 : Résultats du classifieur neuro-génétique paramétrique(CNGP)

✓ **Apprentissage structurel**

Population	génération	Nombre de neurones cachés	CC	Se	Sp	VP	VN	FP	FN
100	100	4	86.34%	100%	82.28%	52	142	31	0

Tableau IV.3 : Résultat du classifieur neuro-génétique structurel(CNGS)

➢ **Interprétation des résultats**

Les classifieurs neuro-géntiques ont donné des meilleures performances particulièrement dans la reconnaissance des cas malins (voir tableaux)

Cette performance est du essentiellement à l'optimisation de la structure et à la modification des poids en plus pour l'apprentissage structurel (2 entrées X3 et X8) ont été éliminé, ces deux derniers représentent respectivement :Uniformity of cell shape et Noarmal Nucleioli

Du point de vu médical, ils sont les moins significatifs (confirmé par les experts du domaine)

IV.2.2 Approche multi agents

Nous avons adopté l'approche distribuée (SMA) pour profiter des contributions des trois modèles (CNC, CNGP, CNGS) représentant chacun un agent auquel nous avons intégré un quatrième agent que nous avons dénommé « AgentContrôleur ».

Au préalable, ces différents agents ont été dénommés comme suit :

1-***Agent Classiffieur*** pour les réseaux de neurones CNC

2-***AgentClassiffeur1*** pour les neuro-génétique paramétrique CNGP

3-***AgentClassifieur2*** pour les neuro-génétique structurel CNGS

L'AgentContrôleur récupère les résultats des trois agents ensuite il calcule ses propres performances afin d'établir le diagnostic final les résultats obtenus sont présentés dans le tableau suivant :

AgentContrôleur		
CC	**Se**	**Sp**
97,5%	**98%**	**97%**

Tableau IV.4 : Résultats AgentContrôleur

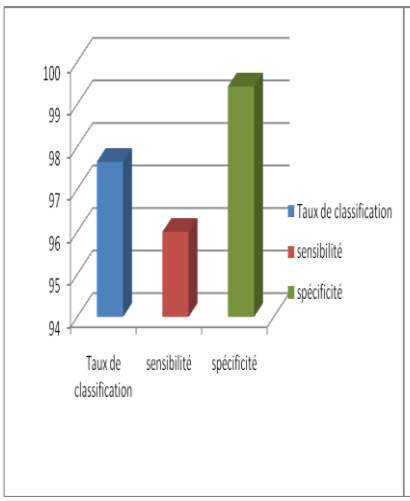
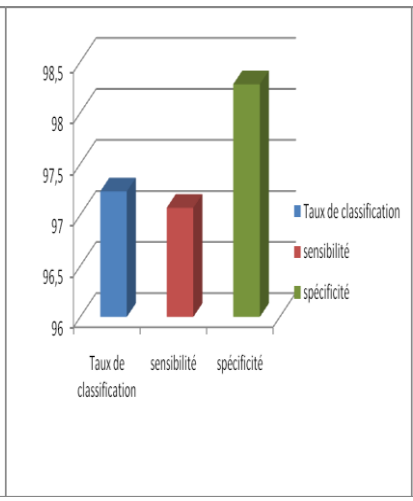
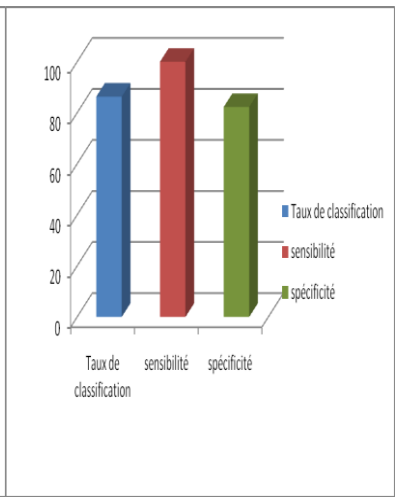


Figure IV.27 : Histrogramme Des performances du CNC



FigureIV.28 : Histogramme des performances du CNGP



FigureIV.29 Histogramme des performances du CNGS

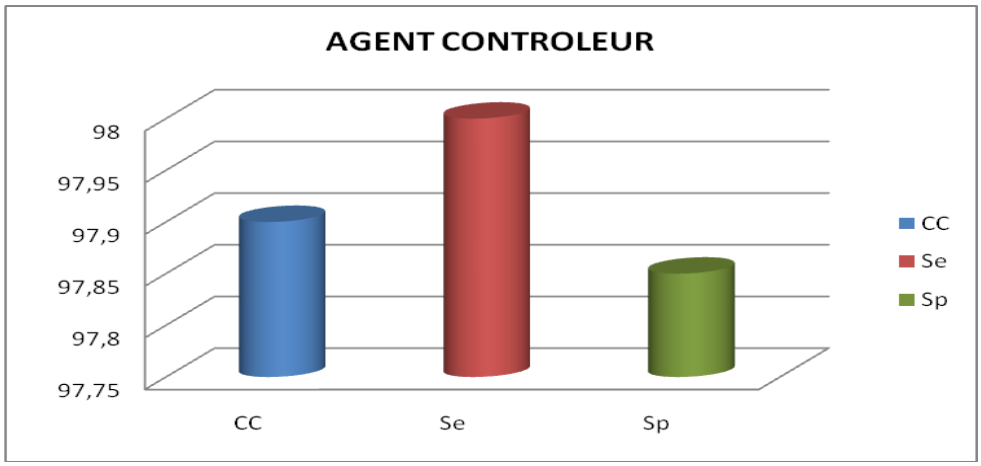


Figure IV.30 : Histrogramme des performances de l'AgentControleur

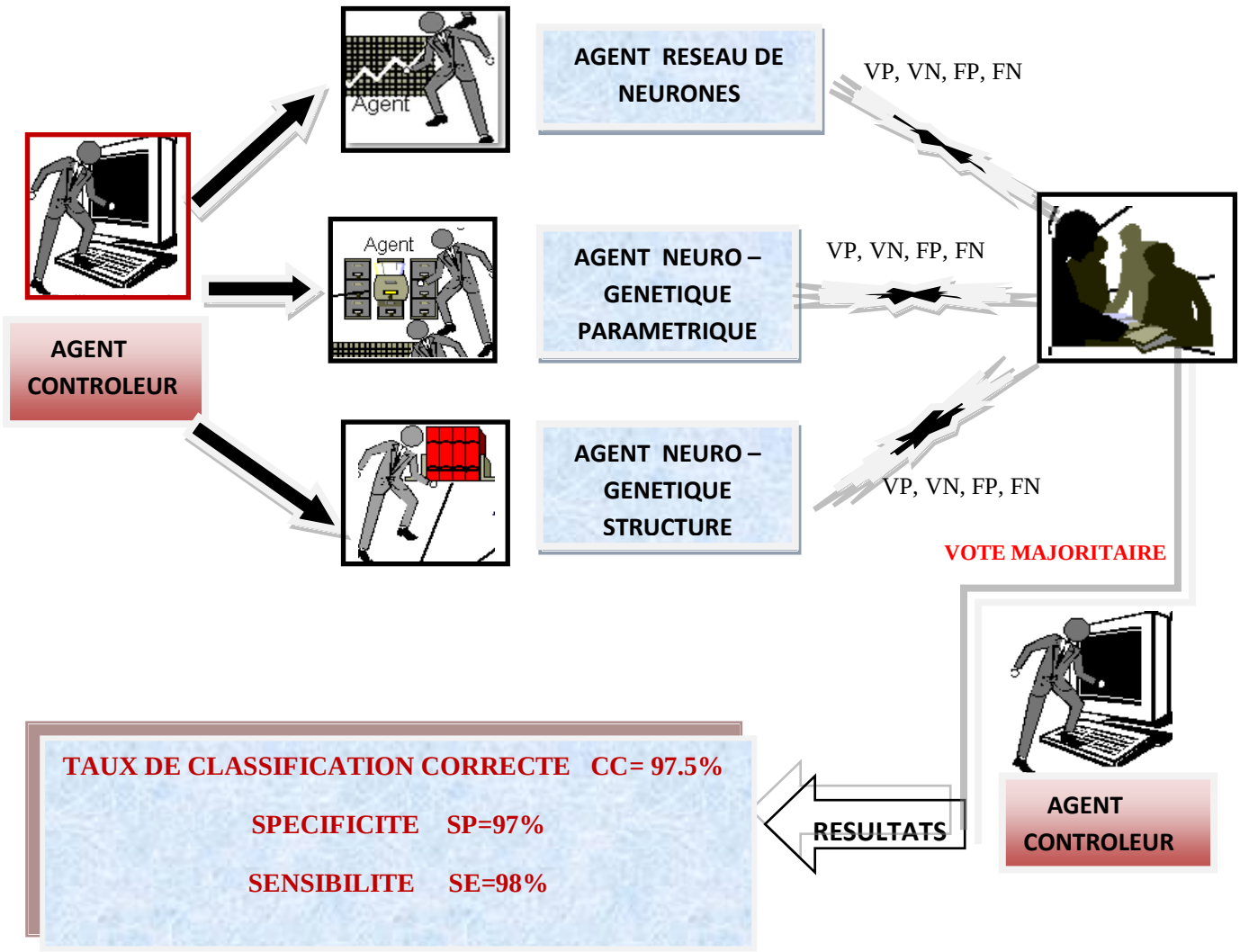


Figure IV.31 : processus de relation agent contrôleur - agents classifieurs

IV.2.3 Conclusion et interprétation :

Nous remarquons que l'AgentContrôleur a bénéficié des meilleurs résultats obtenus par les trois agents (vote majoritaire) pour calculer ses propres performances qui ont augmenté visiblement ses résultats.

IV.2.4 Menu d'utilisation de notre application :

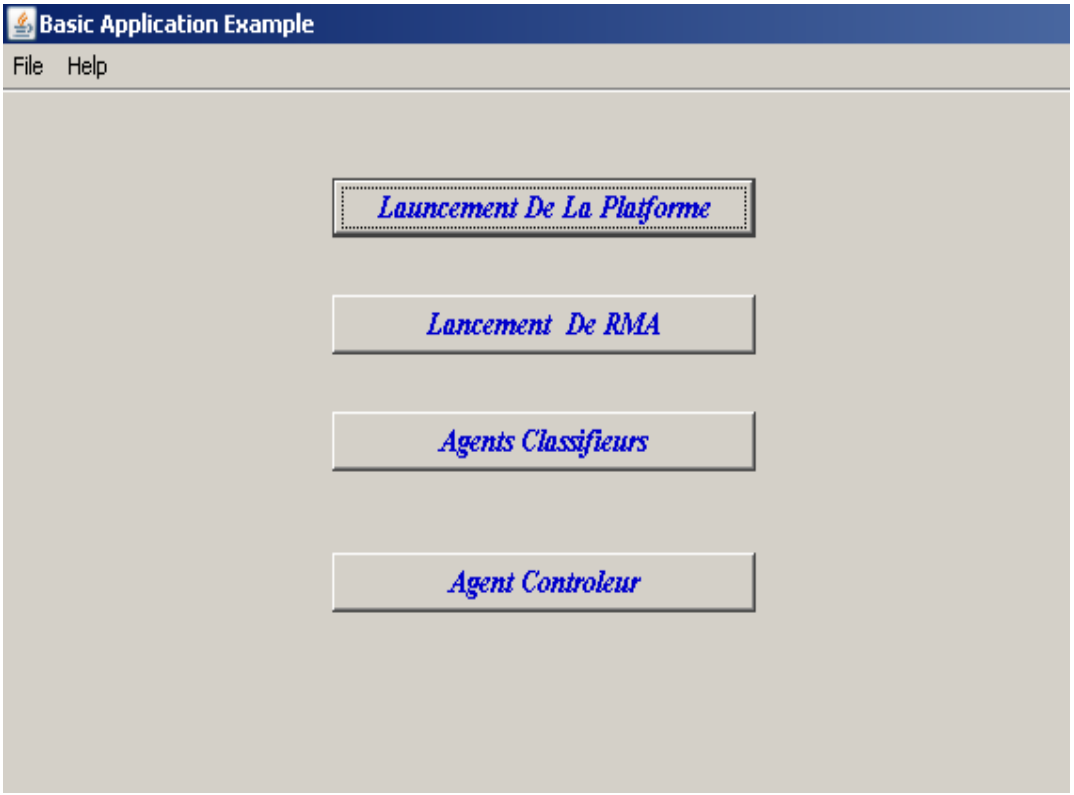


Figure IV.32: Interface crée

Lancement de la plateforme Jade (*launch plateforme) de launch RMA apparition de l'interface*

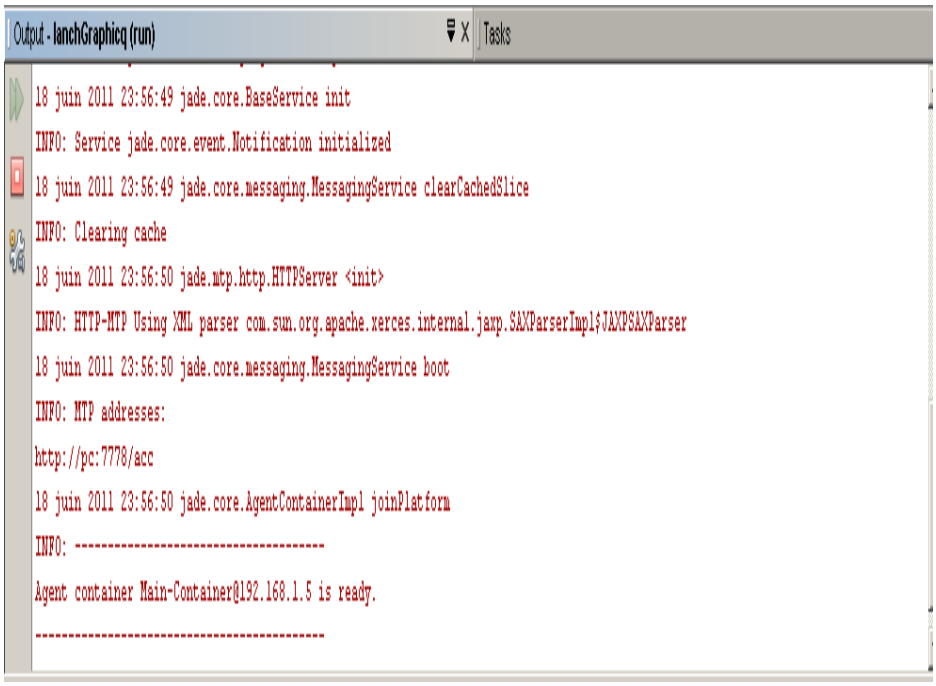


Figure IV.33 :Extrait d'un exemple d'exécution

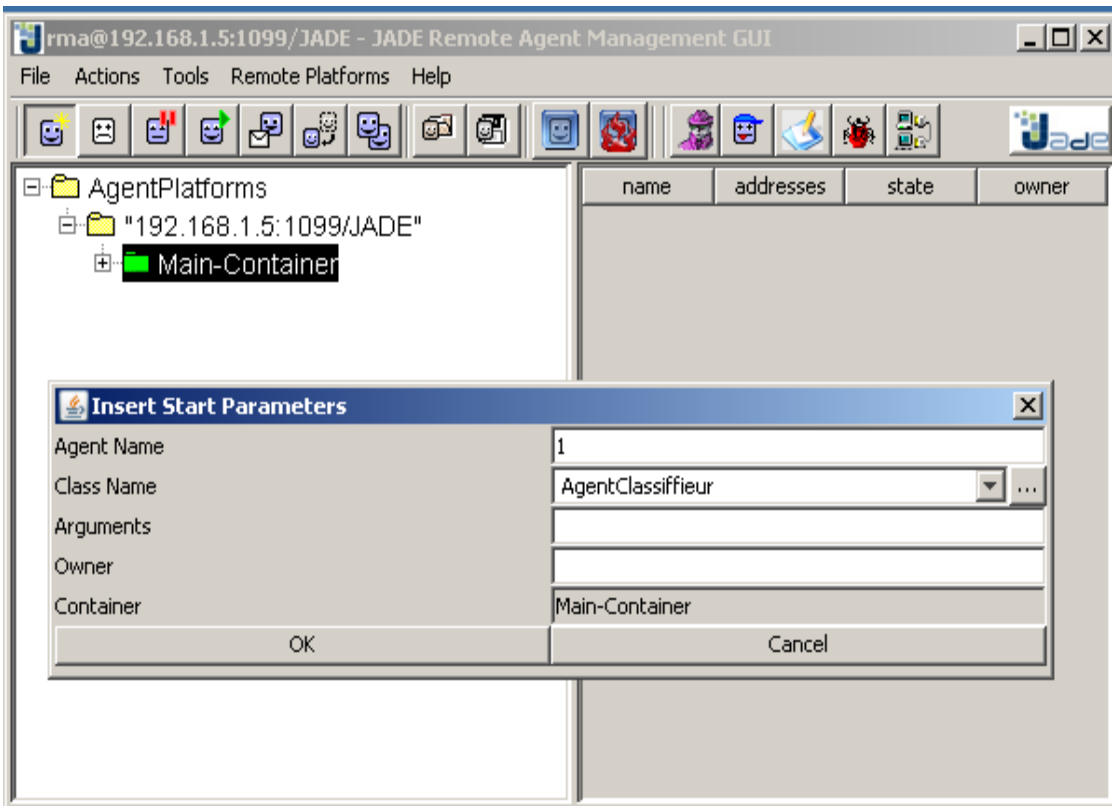


Fig IV.34 : plateforme Jade Création des 3 agents classifieurs

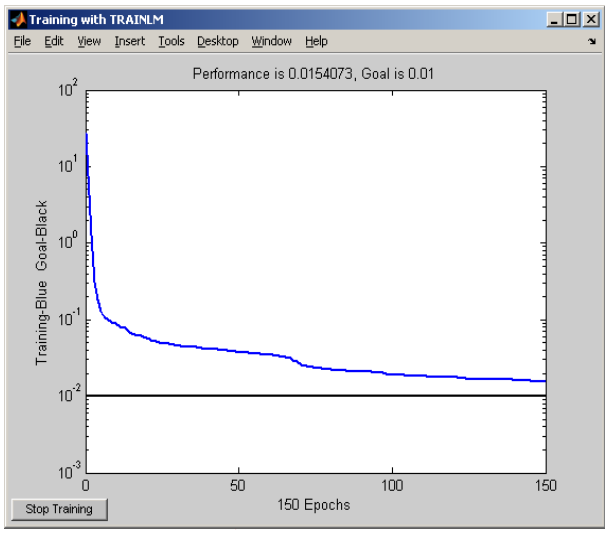


Figure IV.35 :Apprentissage du CNC

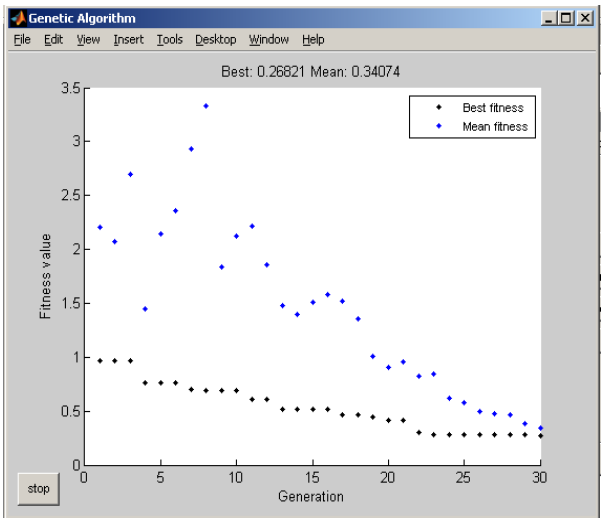


Figure IV.36 :Apprentissage du CNGP

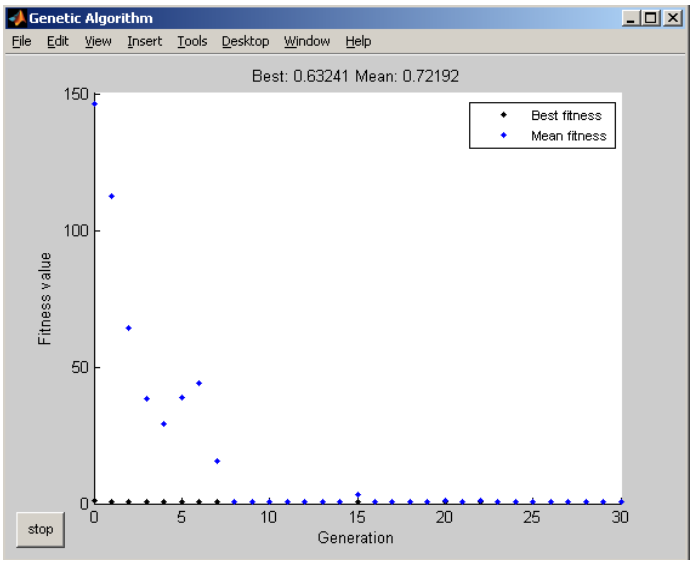


Figure IV.37 : Apprentissage du CNGS

➢Retour sur interface jade pour la création de l'agent Contrôleur : Apparition des résultats (Tableau IV.4)

NB :une autre démarche de raccourcie est possible ,elle consiste à lancer au niveau de l'interface(figure IV.32) **:**

1-La plateforme 2-Agents Classifieurs 3-Agent Contrôleur

CONCLUSION GENERALE

Notre approche multi agents pour la reconnaissance du cancer vise à renforcer le diagnostic d'une manière distribué.

Pour cela nous avons utilisé la base de données universelle WBCD pour évaluer notre modèle.

Nous avons en premier lieu conçu des classifieurs mono agents en utilisant les réseaux de neurones multicouches, leurs hybridations avec les algorithmes génétiques

Nous avons évalué et testé les performances de chaque agent en terme de sensibilité (Se),Spécifité(Sp) et le taux de classification correcte (CC)

En dernier lieu nous avons utilisé un modèle SMA composé de quatre agents : les trois agents conçus auparavant, plus un agent contrôleur.

Les résultats de classifications des données par l'agent ont été très prometteurs.

Notre approche peut se développer davantage par :

-L'élargissement de la base de données

-La multiplication du nombre d'agents (classifieurs)

- L'augmentation de la communication entre différents agents

Nous souhaiterons intégrer la notion d'interpretabilité chez les agents par l'hybridation de leurs techniques de classification avec l'approche floue.

ooooooooooooooooooooooooooooo