الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي
Université Abou Bekr Belkaid جامعة ابي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et
De la Recherche Scientifique

Université Abou Bekr Belkaid -Tlemcen-
Faculté des Sciences-Tidjani Haddam
Département d'Informatique

Mémoire

De fin d'étude
Pour l'obtention du diplôme de Master en Informatique

Option : Système d'information et de connaissance (SIC)

Thème

Classification neuro génétique du diabète

Présentée par :

TEBBAL Karima Kawther

Soutenu en 2011 devant la commission composé de :

Président:	-Mr. LEHSAINI M	
Encadreur:	-Mr. CHIKH M.A	
Examinateurs:	-Mr. BENAMAR A.	- Mr. SMAHI I
	-Mr. BENZIAN Y	- Mr. HADJILA F
	-Mr. BENAZOUZ M	

Année Universitaire : 2010-2011

Table des matières

Liste des figures :

Chapitre I : Présentation du diabète

Chapitre II : Techniques de classification

Chapitre III: Résultats et interprétations

Annexe A

Liste des tableaux :

Chapitre I : Présentation du diabète

Chapitre II : Techniques de classification

Chapitre III: Résultats et interprétations

Liste des abréviations :

ADN : Acide désoxyribonucléique

ANFIS: Adaptative-Network-Based Fuzzy Inference System

AG : Algorithme génétique

CNP : classifieur neuronal probabiliste

CNG : classifieur neuro-génétique

DID : diabète insulinodépendant

DNID : diabète non insulinodépendant

Epai : épaisseur de la peau au niveau du triceps

GDM : diabète gestationnel

Glu : Concentration de glucose dans le plasma

GRNN: Generalized Regression Networks

INS: Taux d'insuline

IMC : Indice de masse corporelle

MNT : Les maladies non-transmissibles

MODY : Maturity Onset Diabetes in the Young

Ngross : Nombre de grossesses

PAD : pression artérielle diastolique

Ped : Fonction pedigree du diabète

PNN : Probabilistic Neural Networks

RNA : Réseau de neurones artificiels

RBR : Réseau à bases radiales

Introduction Générale :

Dans le diagnostic médical, les informations fournies par les patients peuvent inclure des symptômes et des signes redondants liés entre eux en particulier lorsque les patients souffrent de plus d'un type de maladie de la même catégorie. Dans cette situation les médecins trouvent des difficultés pour faire un diagnostic correct. Il est donc nécessaire d'identifier les caractéristiques importantes d'une maladie et cela peut faciliter la tâche des médecins lors d'une consultation donnée.

Aujourd'hui, le diabète et les autres maladies non-transmissibles (MNT) qui partagent les mêmes facteurs de risque représentent un danger majeur pour la santé et le développement humain. On estime que 8 à 14 millions de personnes meurent prématurément chaque année dans les pays en voie de développement en raison de maladies non-transmissibles pouvant faire l'objet de prévention - principalement les maladies cardio-vasculaires, ***le diabète***, les cancers et les maladies respiratoires chroniques. Ces gens meurent trop jeunes à la suite d'une exposition accrue aux facteurs de risque courants pour les maladies non transmissibles. Ces types de maladies sont dus à une alimentation déséquilibrée, la sédentarité, le tabagisme et l'usage nocif de l'alcool.

La population diabétique mondiale ne cesse d'augmenter, en 1998 était de 150 millions, ce chiffre doublera en 2025. Cette épidémie qui concerne surtout le diabète type 2 est liée à plusieurs facteurs dont le vieillissement de la population, les régimes hypercaloriques, l'obésité et les changements de mode de vie dominés par la sédentarité. Il existe une extrême hétérogénéité de la prévalence du diabète d'un pays à l'autre. L'Algérie est en pleine transition épidémiologique et le diabète pose un vrai problème de santé publique par le biais des complications chroniques dominées par les complications cardio-vasculaires, le pied diabétique, l'insuffisance rénale chronique et la rétinopathie. Selon une enquête de l'institut national de santé publique le diabète occupe la quatrième place dans les maladies chroniques non transmissibles.

Actuellement la sensibilisation de la population par l'identification des facteurs de risque qui peuvent être à l'origine du diabète, fait l'objectif des différents acteurs dans le domaine de la santé publique.

Cependant, la reconnaissance et l'identification de ces facteurs repose généralement sur des études faites sur une grande population. Ces études sont regroupées sous forme des bases de données informatisées dans les hôpitaux et les instituts médicaux.

Chaque fois la taille de ses bases de données médicales est grande, l'analyse visuelle et l'exploitation de ses données devient très complexe pour les experts humains. Pour cette raison des techniques dites intelligentes d'extraction et d'analyse de données ont été utilisées pour faire face à cette problématique.

L'utilisation de systèmes de classification pour le diagnostic médical est en développement progressif. Il n'y a aucun doute que l'évaluation des données du patient et les décisions des experts sont les facteurs les plus importants dans le diagnostic. Notons que, les systèmes experts et les différentes techniques d'intelligence artificielle ont prouvé dans les dernières années leurs efficacités d'aider les experts dans le domaine médical.

L'objectif de notre travail consiste à implémenter une approche basée sur les algorithmes génétiques et les réseaux de neurones qui permettra de connaitre si un patient est diabétique ou non diabétique suivant plusieurs descripteurs.

Nous réalisons un couplage entre deux grands domaines à savoir la Médecine et l'Informatique pour le diagnostic du diabète. L'optique de ce couplage est d'améliorer les performances et l'efficacité des systèmes de diagnostic médical.

Notre travail est consacré à l'implémentation de deux modèles intelligents, le premier est basé sur une approche neuronale probabiliste et le deuxième sur une approche neuro génétique pour la reconnaissance automatique du diabète.

Ce mémoire est organisé de la manière suivante :

- Introduction générale
- Chapitre1 : ce chapitre est consacré à la présentation du diabète, ces différentes formes ainsi ces différentes complications et aussi ces traitements et les facteurs de risque associés.
- Chapitre 2 : ce chapitre met en valeur les techniques de classification, nous exposons en 1^{ére} partie, les réseaux de neurones, ces différentes architectures, ces différents types d'apprentissage et ses différentes limites, en seconde partie nous donnons une présentation générale sur les algorithmes génétiques, et enfin l'étude de ces deux techniques nous permet d'exploiter leurs avantages afin de les hybrider dans un même système qui est le modèle neuro-génétique.
- Chapitre 3: nous implémentons et nous comparons deux modèles intelligents pour la reconnaissance automatique du diabète, le premier est un modèle neuronale probabiliste, et le deuxième est un modèle neuro-génétique.

Ce mémoire se termine par une conclusion générale.

Chapitre I

Présentation du diabète

I. présentation du diabète

I.1 Introduction :

Le diabète est parmi les maladies les plus répandues à travers le monde. Actuellement on estime à 150 millions le nombre de personnes atteints de diabète dans le monde. Malgré les efforts de recherche entrepris depuis plusieurs décennies et l'espoir de traitements radicaux, voire préventifs, cette maladie ne bénéficie encore que de traitements substitutifs aux contraintes quotidiennes, rendant active la participation du patient à son traitement.

Dans ce chapitre nous présentons le diabète, ces différentes formes ainsi les différentes complications qui sont souvent associées à cette maladie et aussi ses symptômes et ses traitements.

I.2 Historique :

Le diabète a été reconnu depuis l'Antiquité, et des traitements d'efficacité différente ont été connus dans différentes régions depuis le Moyen Age, et dans la légende pour beaucoup plus longtemps, la pathogenèse du diabète a été entendue expérimentalement depuis environ 1900.

La découverte d'un rôle pour le pancréas dans le diabète est généralement attribuée à **Joseph vonMering** et **Oskar Minkowski**, qui en 1889 a constaté que les chiens dont le pancréas a été enlevée a développé tous les signes et les symptômes du diabète et mourut peu après.

En 1910, Sir **Edward Albert Sharpey-Schafer** a suggéré que les personnes atteintes de diabète ont été déficients en un seul produit chimique qui est normalement produite par le pancréas, il a proposé d'appeler cette substance d'insuline, de la latine insula, ce qui signifie île, en référence aux îlots producteurs d'insuline de Langerhans dans le pancréas.

Le rôle endocrine du pancréas dans le métabolisme, et l'existence de l'insuline, n'a pas été précisée jusqu'en 1921, lorsque Sir Frederick Grant Banting et Charles Herbert Best refait le travail de Von Mering et Minkowski, et est allé plus loin pour démontrer qu'ils pourraient renverser induite diabète chez les chiens en leur donnant un extrait des îlots de Langerhans du pancréas des chiens en bonne santé.

Banting, Best, et ses collègues (et surtout au pharmacien Collip) est allé à purifier l'insuline, une hormone du pancréas de bovins à l' Université de Toronto . Cela a conduit à la disponibilité d'un traitement efficace des injections de l'insuline et le premier patient a été traité en 1922. Pour cela, Banting et directeur du laboratoire MacLeod a reçu le Prix Nobel de physiologie ou médecine en 1923, les deux ont partagé leur prix en argent avec les autres dans l'équipe qui n'ont pas été reconnus, en particulier les meilleures et Collip. **Banting** et **Best** fait le brevet disponibles sans frais et ne tente pas de contrôler la production commerciale.

Insuline production et de la thérapie se propager rapidement à travers le monde, principalement en raison de cette décision. **Banting** est honoré par Journée Mondiale du Diabète qui se tient le jour de son anniversaire, Novembre 14.

La distinction entre ce qui est maintenant connu sous le nom diabète de type 1 et diabète de type 2 a été clairement faite par Sir **Harold Percival (Harry) Himsworth**, et publié en Janvier 1936.

Malgré la disponibilité du traitement, le diabète reste une cause majeure de décès .Par exemple, les statistiques révèlent que la cause spécifique taux de mortalité en 1927 s'élevait à environ 47,7 pour 100.000 habitants en Malte .

En 1980, société de biotechnologie américaine Genentech développé l'insuline humaine. L'insuline est isolée à partir de bactéries génétiquement modifiées (les bactéries contiennent le gène humain de synthèse de l'insuline humaine), qui produisent de grandes quantités d'insuline. L'insuline purifiée est distribué dans les pharmacies pour utilisation par les patients atteints de diabète. **[DIABHIS]**

I.3 Principes du diabète :

Afin de bien comprendre ce qu'est le diabète, il faut savoir comment notre corps prend de l'énergie dans les aliments.

Les êtres vivants ont besoin d'énergie pour vivre. Il suffit de manger pour retrouver de l'énergie. Tous les aliments (le pain, les pommes de terre, les pâtes, etc.) contiennent des sucres, qu'on appelle des glucides. Par contre, pour nourrir les cellules de notre corps, il faut transformer ces glucides en une autre sorte de sucre : le glucose. Cette transformation se produit pendant la digestion. Ainsi, la principale source d'énergie qui nous permet de fonctionner, c'est le glucose. (Voir **Annexe B**)

Le glucose est produit par le foie. Il passe dans le sang et le **taux de glucose** (qu'on appelle aussi **taux de sucre** ou **glycémie**) augmente. À ce moment, le corps envoie un signal à un organe près de l'estomac qui s'appelle le pancréas. Le pancréas produit alors de l'insuline.

L'insuline a beaucoup d'importance. Elle fonctionne un peu comme une clé. C'est elle qui permet au glucose d'entrer dans les cellules pour les nourrir.

Lorsque le glucose entre dans les cellules, le corps reçoit l'énergie dont il a besoin.

Le diabète se développe lorsque le corps manque d'insuline ou ne réussit pas à utiliser celle qu'il produit. Alors, le glucose ne peut plus entrer dans les cellules et s'accumule dans le sang. C'est ce qu'on appelle faire de l'hyperglycémie. **[DIAB6]**

I.3.1 Définition du diabète :

Le diabète, ou diabète sucré, est une maladie chronique provoquée par un trouble du métabolisme du glucose qui perturbe le stockage et l'utilisation par l'organisme de ce carburant nécessaire à son énergie et caractérisée par un taux anormalement élevé de sucre dans le sang et les urines. La glycémie normale (taux de sucre) à jeun varie de 0,8 g à 1 g par litre de sang. Ce trouble résulte soit d'un défaut, partiel ou complet, du pancréas à synthétiser l'insuline pour absorber le glucose. Comme il est mal absorbé par les cellules, le glucose s'accumule dans le sang et cause l'hyperglycémie (une augmentation de la concentration du sang en glucose). Les cellules étant privées de leur principale source d'énergie, il s'ensuit forcément des conséquences physiologiques importantes. **[BIODIAB]**

I.4 Les différentes formes de diabète :

I.4.1 Les diabètes primaires : type 1, type 2, gestationnel :

La plupart des cas de diabète sucré se répartissent en trois grandes catégories : type 1, type 2 et diabète gestationnel.

I.4.1.1 Diabète de type 1 :

Le diabète de type 1 appelé autrefois diabète insulinodépendant (DID) (ou encore diabète juvénile), appelé aussi diabète maigre, touche environ 10 à 15% des diabétiques et apparaît en général chez les personnes âgées de moins de 35 ans, particulièrement pendant l'enfance et l'adolescence. Dans ce cas, le diabète est une forme de maladie auto-immune, c'est-à-dire une maladie dans laquelle le corps stimule ses mécanismes de défense naturelle contre lui-même. Il est dû à une destruction, à plus de 90%, des cellules ß (béta) du pancréas fabriquant l'insuline, hormone indispensable à la vie. Cette forme de diabète peut affecter l'individu dès son plus jeune âge et nécessite obligatoirement des injections pluriquotidiennes d'insuline. Il existe une prédisposition génétique à développer le diabète de type 1. Le risque d'avoir un enfant diabétique pour un parent diabétique est de 10 %.

I.4.1.2 Diabète de type 2 :

Le diabète de type 2 appelé autrefois non insulinodépendant (DNID) (ou diabète de l'âge mûr), connu aussi sous le nom de diabète gras, ce diabète survient classiquement chez l'adulte de plus de 40 ans présentant, dans 80 % des cas, une obésité ou du moins un excès pondéral. Il est quelque fois précédé du diabète de type 1.

Au début de la maladie, la production d'insuline par le pancréas est normale. Mais, les cellules de l'organisme chargées de capter et d'utiliser le glucose deviennent insensibles à l'insuline, d'où une augmentation de la glycémie. **[LPATHO]**

I.4.1.3 Comparaison entre diabète de types 1 et 2 :

D'entité	Diabète de type 1	Diabète de type 2
Apparition	Brusque	Progressive
L'âge de début	Tout âge (Surtout des jeunes)	Principalement chez les adultes
Conseil d'habitus	Mince ou normale	Souvent obèses
L'acidocétose	Commune	Rare
Les auto-anticorps	Habituellement présente	Absent
l'insuline endogène	Faible ou nulle	Normal, diminué ou augmentée
Concordance dans les jumeaux identiques	50%	90%
Prévalence	Moins fréquents	Plus fréquents - 90 à 95% des diabétiques américains

Tableau I.1 : Comparaison du diabète de types 1 et 2

I.4.1.4 Diabète gestationnel :

Le diabète gestationnel (GDM) ou diabète de grossesse est un diabète qui apparaît durant la grossesse, habituellement pendant le 2^{e} ou le 3^{e} trimestre. Les médecins posent aussi un diagnostic de diabète gestationnel lorsqu'une intolérance au glucose (état pré diabétique) est détectée chez une femme enceinte. Autrement dit, le diabète gestationnel n'est pas à tout coup un diabète franc, mais dans tous les cas, la glycémie (ou taux de « sucre » dans le sang) est supérieure à la normale.
Parfois, le diabète était présent avant la grossesse, mais n'avait pas encore été dépisté. Un test de glycémie est souvent proposé aux femmes enceintes en début ou en milieu de grossesse, selon leur risque d'être atteintes du diabète gestationnel.

le diabète gestationnel est une préoccupation croissante : il touche maintenant environ 7 % des femmes enceintes. Le taux est beaucoup plus élevé dans les populations autochtones : 13 %, en moyenne. **[DIABGDM]**

I.4.2 Les diabètes secondaires :

Les autres formes de diabète sont beaucoup plus rares, représentant chacune quelques pourcent des cas.

- les diabètes de types MODY, ont la particularité d'être génétiquement déterminés, selon un mode de transmission autosomique dominant: dans les familles porteuses, atteinte d'un individu sur 2, à toutes les générations. Le début en est habituellement précoce (néonatal parfois, avant 25 ans en général), et le plus souvent ils réalisent des diabètes non insulinodépendants.
- Les diabètes secondaires à des maladies du pancréas (pancréatite chronique, cancer du pancréas, mucoviscidose, hémochromatose, chirurgie du pancréas).
- Les diabètes secondaires à des maladies endocrines, dont le syndrome de Cushing, l'acromégalie, le phéochromocytome, l'hyperthyroïdie, l'adénome de Conn, etc.

- ❖ Les diabètes secondaires à des maladies du foie, cirrhose, quelle qu'en soit la cause, mais plus particulièrement dans le contexte de l'infection par le virus C de l'hépatite (hépatite virale C), ou l'hémochromatose.
- ❖ les diabètes secondaires à des mutations de l'ADN mitochondrial (associé à une surdité de perception et caractérisé par une hérédité maternelle) : syndrome de Bal linger-Wallace.
- ❖ le diabète lipoatrophique : Lipodystrophie congénitale de Berardinelli-Seip, caractérisé par la disparition du tissu adipeux, avec insulino-résistance majeure, hyperlipidémie et stéatose hépatique ;
- ❖ Les diabètes associés à des médicaments, en particulier les corticoïdes, les diurétiques, les antipsychotiques (risperdal), les immunosuppresseurs de la famille des inhibiteurs de la calcineurine, etc. **[WIKI2]**

I.5 Les symptômes du diabète :

Les symptômes ne se présentent pas tous de la même manière ni avec la même intensité. Qu'il s'agisse du type 1, du type 2 ou du diabète de grossesse, une consultation avec le médecin s'impose. Les symptômes sont :

- fatigue, somnolence
- augmentation du volume des urines
- soif intense
- faim exagérée
- amaigrissement
- vision embrouillée
- cicatrisation lente
- infection des organes génitaux
- picotements aux doigts ou aux pieds
- changement de caractère

Il est à noter que, parfois, les symptômes ne sont pas apparents. Le diabète est une maladie grave. Il peut avoir un impact considérable sur la qualité de vie des personnes qui vivent avec cette maladie. **[DIABSY]**

I.6 Les complications liées au diabète :

Le diabète s'attaque aux veines et aux artères, surtout aux plus petites. Le sang y circule plus difficilement et ne peut plus bien nourrir les cellules qui les entourent. Les muscles et les nerfs peuvent être abîmés. De plus, comme le sang circule moins bien, le corps peut moins combattre les infections. C'est pour ces raisons que les complications liées au diabète sont nombreuses et souvent graves.

Le diabète peut affecter vos yeux, entraîner des problèmes cardiaques ou vasculaires, perturber le fonctionnement de vos reins et endommager votre peau, notamment au niveau des pieds.

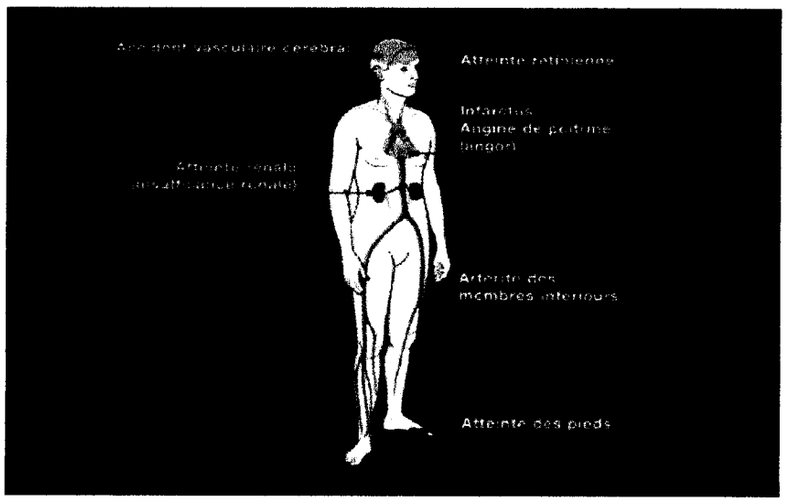


Figure I.1 : Les complications du diabète [COMPDIA]

I.7 Les traitements du diabète :

A ce jour, les médecins n'ont pas encore trouvés de cure permettant de guérir le diabète, mais une médication adéquate, un bon régime alimentaire et quelques modifications au mode de vie peuvent permettre aux personnes diabétiques de mener une vie pratiquement normale tout en évitant à long terme les problèmes et les complications souvent associés à cette maladie. **[BIODIAB]**

Il faut connaitre que la seule façon de savoir si une personne est atteinte de diabète, c'est de faire vérifier son taux de sucre par une prise de sang.

I.7.1 Contrôler la glycémique :

Avoir un bon équilibre glycémique permet :

- Au quotidien de limiter la fréquence et la sévérité des hypoglycémies.
- De réduire les hyperglycémies et donc les risques de complications à moyen et long termes.

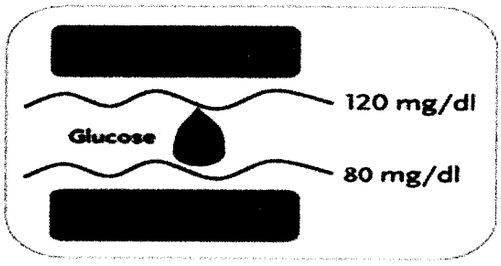


Figure I.2 : Equilibre glycémique [LMSGY]

➢ **la glycémie :**

La glycémie est une mesure de la concentration du glucose dans le sang. Les personnes diabétiques doivent surveiller leur glycémie de prés afin d'ajuster leur médication (en fonction de l'alimentation, de l'exercice physique, du stress, etc.) et de maintenir en tout temps une glycémie aussi près que possible de la normale. Le contrôle de la glycémie est d'autant plus important qu'il permet de réduire ou de prévenir les complications du diabète. **[BIODIAB]**

➢ **L'hyperglycémie et L'hypoglycémie :**

L'hyperglycémie et l'hypoglycémie sont deux problèmes qui sont souvent liés au diabète.

- **Hyperglycémie :** une augmentation de la concentration de glucose dans le sang : lorsqu'à jeun, la glycémie est supérieure ou égale à 7 m mol/l (120 mg/dl) ou qu'une à deux heures après un repas, elle s'élève à 11 m mol/l (200 mg/dl) ou plus. les symptômes suivants peuvent apparaître : une élimination excessive d'urine, faiblesse, douleurs à l'estomac sensation de douleur générale, respiration difficile et bruyante, une soif et une faim accrues. **[DIAB15]**
- **Hypoglycémie :** une diminution de la concentration de glucose dans le sang : lorsque la glycémie s'abaisse en dessous de 4 m mol/l (80 mg/dl). Les symptômes sont les suivants : tremblements, sueurs, étourdissements, palpitations, fatigue, bâillements, vue embrouillée, etc. **[DIAB15] [BIODIAB]**

I.7.2 les traitements de chaque type de diabète :

Ce qui concerne les traitements, il est important de savoir si on a affaire à un diabète de type 1 ou de type 2 ou d'autre type. En effet, ils peuvent changer selon le type de diabète. **[DIAB21]**

❖ **diabète de type 1 :**

Pour le diabète de type 1, il y a un seul traitement possible : **l'injection d'insuline**, par voie sous-cutanée, 1 à 4 fois par jour. On utilise une insuline artificielle, fabriquée par génie génétique, qui a exactement la même composition que l'insuline humaine. Pour être autonomes, les personnes souffrant de diabète insulinodépendant, y compris les enfants, apprennent à pratiquer eux-mêmes les injections. Les diabétiques de type 1 sont ce que l'on appelle **insulinodépendants**, c'est à-dire qu'ils ont besoin d'insuline pour vivre.

Figure I.3 : Personne se faisant une injection d'insuline à l'aide d'un stylo.

Ce système d'injection, simple à mettre en œuvre, a permis d'améliorer la qualité de vie des diabétiques.

Il existe différentes méthodes : utilisation d'une seringue classique, l'insuline étant pompée dans un flacon qui doit être conservé au frais, stylo injecteur muni d'une réserve d'insuline, plus maniable. Le patient peut également être équipé d'un dispositif portable (à la taille ou en bandoulière), relié à une seringue-réservoir d'insuline et muni d'un mécanisme pousse-seringue (pompe à insuline). Ce système est relié à une aiguille implantée sous la peau. Une dernière méthode, encore utilisée à titre expérimental, consiste à implanter sous la peau une pompe à insuline automatique, à débit variable et programmable. Cette pompe est remplie tous les mois ou tous les trois mois, avec une seringue. **[SUCREDIA]**

- **Diabète de type 2 :**

Pour le diabète de type 2, il y a plus de possibilités. En effet, les gens atteints de cette maladie sont **non-insulinodépendants**, c'est-à-dire qu'ils n'ont pas nécessairement besoin d'insuline pour vivre. En fait, il existe un certain nombre de médicaments qui peuvent permettre de contrôler la maladie. Ces médicaments sont de quatre types :

- Ceux qui stimulent la production de l'insuline qui existe encore dans le pancréas.
- Ceux qui aident les cellules à mieux utiliser l'insuline disponible dans le sang.
- Ceux qui empêchent le glucose d'entrer dans le sang pendant la digestion.
- Ceux qui empêchent le foie de fabriquer du glucose.

Ces médicaments peuvent suffire à contrôler le diabète de type 2 s'ils sont associés à de bonnes mesures de prévention. Mais après une dizaine d'années, il arrive que la maladie devienne plus difficile à contrôler. À ce moment, l'injection d'insuline peut devenir nécessaire.

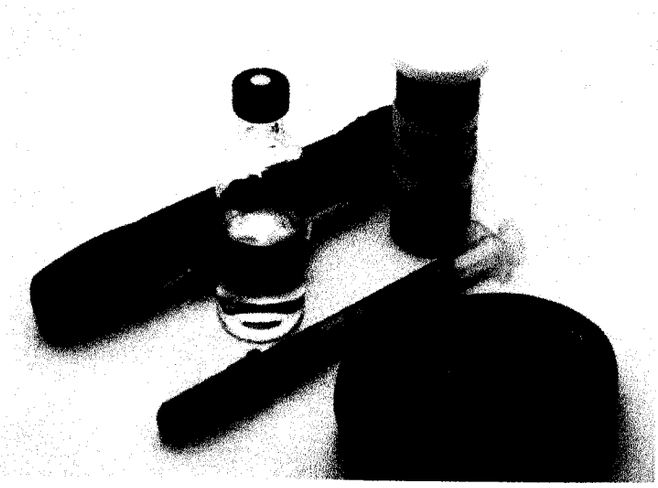


Figure I.4 : les produits qui servent à traiter le diabète [DIAB24]

I.7.3 Mode de vie du diabétique :

En dehors de la médication, les diabétiques ont grand intérêt à établir un plan d'alimentation et à adopter un bon programme d'exercices physiques. En effet, ces interventions non médicamenteuses peuvent permettre de diminuer le dosage de la médication et de prévenir certaines complications.

I.7.3.1 Le régime alimentaire :

Le diabétique doit particulièrement veiller à son alimentation qui devra être régulière, équilibrée et peu calorique. Il devra consommer 50% de sa ration quotidienne sous forme de glucides lents : pain complet, pâtes, riz, légumes secs. Il choisira des fruits crus plutôt que sous forme de compotes. Le reste de l'alimentation comprendra 30 % de lipides (graisses) et 20% de protéines. En effet, cette maladie est directement liée à ce qu'on mange. C'est pourquoi il faut éviter les aliments trop gras ou trop sucrés. **[LPATHO]**

Pour les gens atteints du diabète de type 1, on recommande de répartir la consommation de glucides tout au long de la journée (trois repas et trois collations, par exemple). On suggère de manger plus souvent, mais en petite quantité. Il est ainsi plus facile de stabiliser le taux de glucose dans le sang.

Pour les gens qui ont un diabète de type 2, c'est un peu le même genre de recommandation. Toutefois, comme l'obésité est l'une des premières causes de cette maladie (8 personnes sur 10 qui ont ce type de diabète sont obèses), il va de soi que les gens qui en souffrent doivent également perdre du poids. **[DIAB28]**

I.7.3.2 L'exercice physique :

La surcharge pondérale et le manque d'exercice physique constituent de réels risques d'aggravation et de complications du diabète. Dans bien des cas, il suffit de surveiller son alimentation et de perdre du poids tout en faisant régulièrement de l'exercice pour tenir la maladie en échec et prévenir les problèmes associés, surtout dans le cas du diabète de type 2, ou l'obésité est souvent concomitante. Il est particulièrement important de pratiquer des exercices cardiovasculaires d'intensité modérée, selon le goût : la marche, le tennis, la bicyclette, la natation, etc. Les spécialistes de la clinique Mayo recommandent une séance quotidienne d'au moins 30 minutes, en plus d'ajouter à son programme des exercices d'étirement et de musculation avec poids et haltères.

I.8 Conclusion :

De nos jours, le diabète est toujours une maladie grave. Même si la médecine permet de la contrôler de mieux en mieux, cette maladie touche de plus en plus de monde. Pour améliorer les choses, il faut que les gens se prennent en main, ils peuvent améliorer leur propre alimentation et faire un peu plus d'exercice.

Chapitre II

Techniques de classification

II. techniques de classification

II.1 Les réseaux de neurones

II.1.1 Introduction :

Tout d'abord, ce que l'on désigne habituellement par « réseau de neurones »est en fait un réseau de neurones artificiels (RNA) basé sur un modèle simplifié de neurone. Ce modèle permet certaines fonctions du cerveau, comme la mémorisation associative, l'apprentissage par l'exemple, le travail en parallèle, mais le neurone artificiel est loin de posséder toutes les capacités du neurone biologique. Les réseaux de neurones biologiques sont ainsi beaucoup plus compliqués que les modèles mathématiques et informatiques. **[RN REC]**

Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de générer des classifications rapides, et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme perceptif indépendant des idées propres de l'implémenter, et fournissant des informations d'entrée au raisonnement logique formel. **[RN WIKI]**

L'idée de l'apprentissage artificiel est de s'inspirer de l'apprentissage naturel du cerveau de l'être vivants pour développer des modèles intelligents et autonome. Il existe deux courants de recherche sur les réseaux de neurones : un premier est motivé par l'étude et la modélisation des phénomènes naturels d'apprentissage à l'aide de réseaux de neurones où la pertinence biologique est importante ; un second est motivé par l'obtention d'algorithmes efficaces ne se préoccupant pas de la pertinence biologique (c'est le point le plus utilisé**). [CHIKH 2010]**

II.1.2 Historique :

Les premiers biophysiciens de Chicago, **MacCulloch** et **Pitts**, inventaient en 1943 le premier neurone formel qui portait leurs noms (neurone de **MacCulloch-Pitts**).Quelques années plus tard, en 1949, **Hebb** propose une formulation du mécanisme d'apprentissage, sous la forme d'une règle de modification des connexions synaptiques (règle de **Hebb**).

Le premier réseau de neurones artificiels apparait en 1958, grâce aux travaux de **Rosenblatt** qui conçoit le fameux Perceptron. Le Perceptron est inspiré du système visuel (en termes d'architecture neurobiologique) et possède une couche de neurones d'entrée ("perceptive") ainsi qu'une couche de neurones de sortie ("décisionnelle"). Ce réseau parvient à apprendre à identifier des formes simples et à calculer certaines fonctions logiques. Il constitue donc le premier système artificiel présentant la capacité d'apprendre par l'expérience. Malgré tout l'enthousiasme que soulève le travail de **Rosenblatt** dans le début des années 60, la fin de cette décennie sera marquée en 1969, par une critique violente du Perceptron par **Minsky** et **Papert**. Ils montrent dans un livre «Perceptrons » toutes les limites de ce modèle.

En 1982 le génie **Hopfield** démontre tout l'intérêt d'utiliser des réseaux récurrents (dits "feed-back") pour la compréhension et la modélisation des processus mnésiques.

En parallèle des travaux de **Hopfield, Werbos** conçoit son algorithme de rétro propagation de l'erreur, qui offre un mécanisme d'apprentissage pour les réseaux multicouches de type perceptron (appelés MLP pour Multi-layer Perceptron), fournissant ainsi un moyen simple d'entraîner les neurones des couches cachées. Cet algorithme de "back-propagation" ne sera pourtant popularisé qu'en 1986 par **Rumelhart**. **[RN ing][RN course]**

II.1.3 Neurone biologique :

II.1.3.1 Structure d'un neurone :

On pense que le système nerveux compte plus de 1000 milliards de neurones interconnectés. Bien que les neurones ne soient pas tous identiques, leur forme et certaines caractéristiques permettent de les répartir en quelques grandes classes. En effet, il est aussi important de savoir, que les neurones n'ont pas tous un comportement similaire en fonction de leur position dans le cerveau. Avant de rentrer plus en avant dans les détails, examinons un neurone.

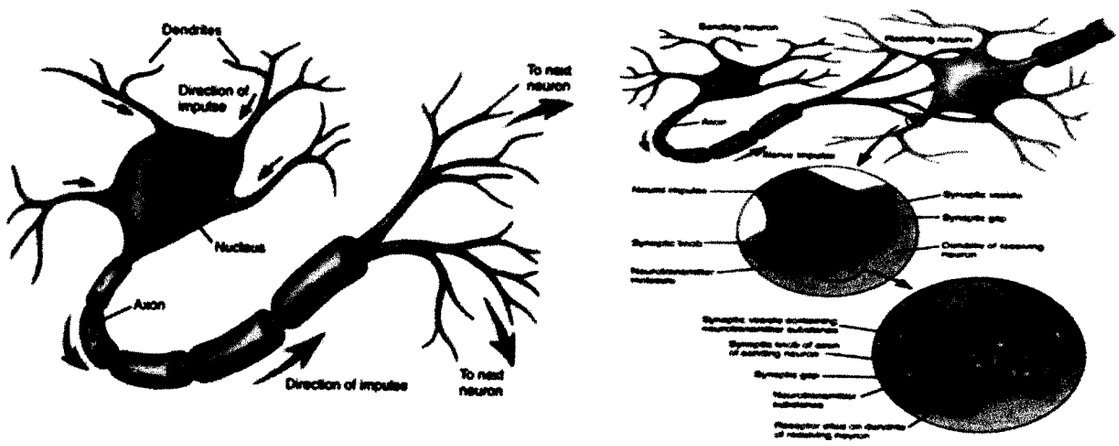


Figure II.5 : Schéma de principe d'un neurone biologique [SCHEMA]

On peut le décomposer en trois régions principales:

- **Le corps cellulaire :** Il contient le noyau du neurone ainsi que la machine biochimique nécessaire à la synthèse des enzymes. Ce corps cellulaire de forme sphérique ou pyramidale contient aussi les autres molécules essentielles à la vie de la cellule. Sa taille est de quelques microns de diamètre.
- **Les dendrites :** Ce sont de fines extensions tubulaires qui se ramifient autour du neurone et forment une sorte de vaste arborescence. Les signaux envoyés au neurone sont captés par les dendrites. Leur taille est de quelques dizaines de microns de longueur.

- **L'axone:** C'est le long de l'axone que les signaux partent du neurone. Contrairement aux dendrites qui se ramifient autour du neurone, l'axone est plus long et se ramifie à son extrémité ou il se connecte aux dendrites des autres neurones. Sa taille peut varier entre quelques millimètres à plusieurs mètres.
- **Les Synapse** : Une synapse est une jonction entre deux neurones, et généralement entre l'axone d'un neurone et une dendrite d'un autre. **[RN COUR]**

II.1.3.2 Fonctionnement :

Au point de vu fonctionnel, il faut considérer le neurone comme une entité polarisée, c'est-à-dire que l'information ne se transmet que dans un seul sens : des dendrites vers l'axone.

Le neurone va donc recevoir des informations, venant d'autres neurones, grâce à ses dendrites. Il va ensuite y avoir sommation, au niveau du corps cellulaire, de toutes ces informations et via un potentiel d'action (un signal électrique) le résultat de l'analyse va transiter le long de l'axone jusqu'aux terminaisons synaptiques.

A cet endroit, lors de l'arrivée du signal, des vésicules synaptiques vont venir fusionner avec la membrane cellulaire, ce qui va permettre la libération des neurotransmetteurs (médiateurs chimiques) dans la fente synaptique. Le signal électrique ne pouvant pas passer la synapse (dans le cas d'une synapse chimique), les neurotransmetteurs permettent donc le passage des informations, d'un neurone à un autre.

Les neurotransmetteurs excitent (neurotransmetteurs excitateurs) ou inhibent (neurotransmetteurs inhibiteurs) le neurone suivant et peuvent ainsi générer ou interdire la propagation d'un nouvel influx nerveux. En effet, au niveau post-synaptique, sur la membrane dendritique, se trouvent des récepteurs pour les neurotransmetteurs. Suivant le type de neurotransmetteur et le type des récepteurs, l'excitabilité du neurone suivant va augmenter ou diminuer, ce qui fera se propager ou non l'information.

Les synapses possèdent une sorte de «mémoire» qui leur permet d'ajuster leur fonctionnement. En fonction de leur «histoire», c'est-à-dire de leur activation répétée ou non entre deux neurones, les connexions synaptiques vont donc se modifier. Ainsi, la synapse va faciliter ou non le passage des influx nerveux. Cette plasticité est à l'origine des mécanismes d'apprentissage. **[RN INDEX]**

II.1.4 Neurone formel : [RN COUR]

Le premier neurone formel est apparu en 1943. On le doit à Mac Culloch et Pitts. Voici un schéma de leur modèle de neurone formel :

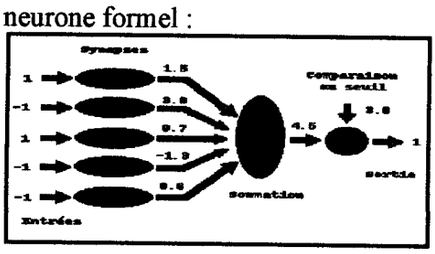


Figure II.6 : Le neurone formel

Le neurone formel est donc une modélisation mathématique qui reprend les principes du fonctionnement du neurone biologique, en particulier la sommation des entrées. Sachant qu'au niveau biologique, les synapses n'ont pas toutes la même «valeur» (les connexions entre les neurones étant plus ou moins fortes), les auteurs ont donc créé un algorithme qui pondère la somme de ses entrées par des poids synaptiques (coefficients de pondération). De plus, les 1 et les -1 en entrée sont là pour figurer une synapse excitatrice ou inhibitrice.

II.1.4.1 Interprétation mathématique :

D'un point de vue mathématique, le neurone formel peut être représenté de la manière suivante:

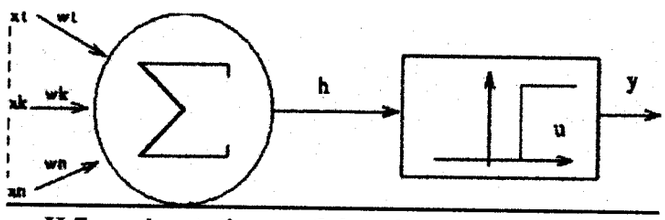


Figure II.7 : présentation mathématique d'un neurone formel

Pour un nombre compris entre j (=1) et un nombre quelconque n , le neurone formel va calculer la somme de ses entrées (x1, ... , xn), pondérées par les poids synaptiques (w1, ... , wn), et la comparer à son seuil téta. Si le résultat est supérieur au seuil, alors la valeur renvoyée est 1, sinon la valeur renvoyée est 0.

D'où la formule (1) (avec f = fonction seuil): $$y = f(\sum_{j=1}^{n} w_j x_j - \theta)$$

II.1.4.2 Principe de fonctionnement :

Le neurone se présente comme une cellule possédant plusieurs liens d'entrée dits "liens synaptiques" et un lien de sortie. Le lien de sortie d'une cellule peut être connecté à un ou plusieurs liens d'entrée d'autres cellules ; il est porteur d'une valeur qu'il transmet à ces liens d'entrée, cette valeur est déterminée à partir des valeurs des entrées propres de sa cellule.

Chaque neurone artificiel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance des neurones amont. A chacune de ces entrées est associé un poids **w** (abréviation de **weight** en anglais) représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones avals.

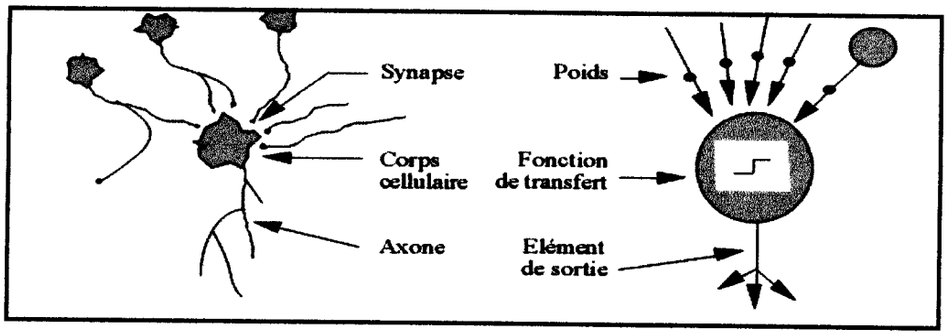


Figure II.8 : Mise en correspondance neurone biologique et neurone artificiel

On pourra résumer cette modélisation par le tableau II.2, qui montre la transition entre le neurone biologique et le neurone formel.

Neurone biologique	Neurone artificiel
Dendrite	Signal d'entrée
Synapses	Poids de connexions
Somma	Fonction d'activation
Axones	Signal de sortie

Tableau II.2 : Analogie entre le neurone biologique et le neurone artificiel [TITI]

II.1.4.3 Fonctions d'activation :

Dans sa première version, le neurone formel était donc implémenté avec une fonction à seuil (a), mais de nombreuses versions existent. Ainsi le neurone de McCulloch et Pitts a été généralisé de différentes manières, en choisissant d'autres fonctions d'activations, comme les fonctions linéaires par morceaux (b), des sigmoïdes (c) ou des gaussiennes (d) par exemples.

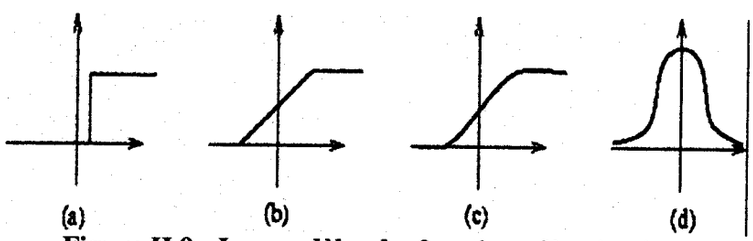


Figure II.9 : Les modèles de fonctions d'activation.

II.1.5 Les différentes architectures des réseaux de neurones :

Les RNA peuvent être vus comme des graphes orientés dans lesquels les neurones formels sont les sommets, et les arcs (pondérés) sont les connexions entre sorties de neurones et entrées de neurones. En fonction du type de connexions (architectures), les RNA sont regroupés en deux grandes catégories (Voir la Figure II.10).

1. réseaux « feedforward », qui sont des graphes acycliques ;
2. réseaux récurrents ou « feedback », qui sont des graphes avec circuits du fait de la présence d'arcs de rétroaction.

Dans la famille la plus répandue des réseaux non-bouclés, celle des PMC, les cellules sont organisées en couches, avec des connexions inter-couches (mais pas de connexions intra-couches). D'une manière générale, différents types de connexions induisent différents types de comportements. Les réseaux de type non bouclé sont statiques, dans le sens où ils ne produisent qu'une seule réponse en fonction des entrées fournies. Ils sont sans mémoire, puisque les réponses fournies ne prennent pas en compte l'état antérieur du réseau. Au contraire, les réseaux récurrents sont dynamiques et dépendants de l'état antérieur. **[BENDI 2008]**

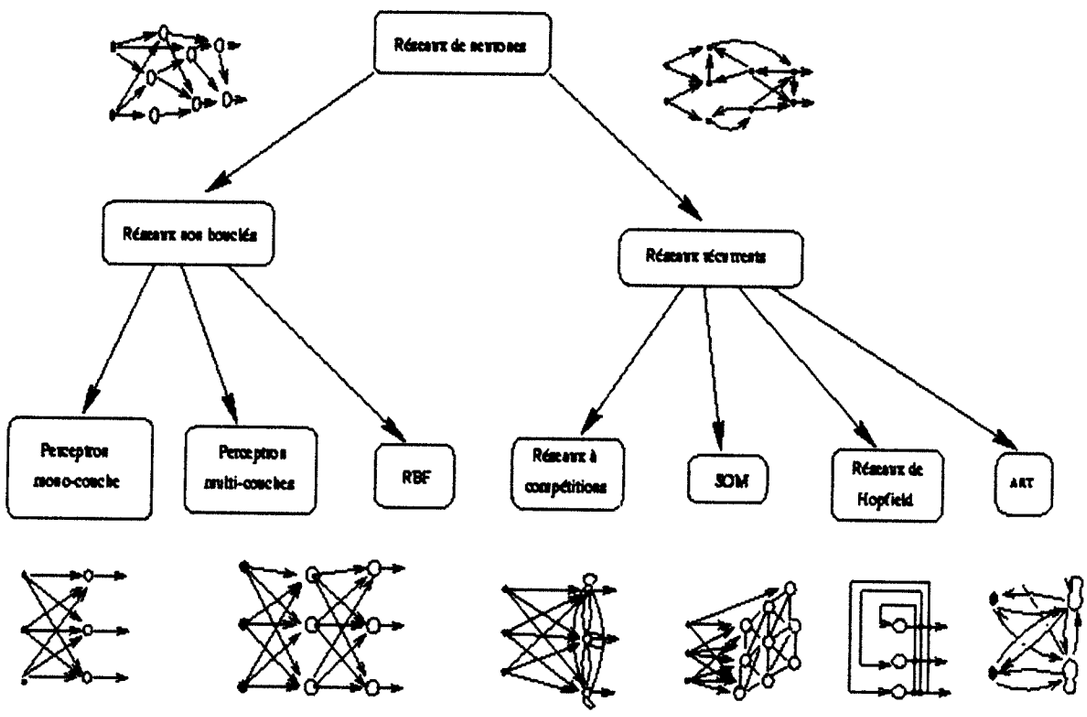


Figure II.10 : les différentes topologies de RNA

II.1.5.1 LES RESEAUX "FEED-FORWARD" :

Appelés aussi "réseaux de type Perceptron", ce sont des réseaux dans lesquels l'information se propage de couche en couche sans retour en arrière possible.

II.1.5.1.1 Les Perceptrons :

✓ **Le perceptron monocouche :**

Historiquement le premier RNA, c'est le Perceptron de Rosenblatt. C'est un réseau simple, puisque il ne se compose que d'une couche d'entrée et d'une couche de sortie. Il est calqué, à la base, sur le système visuel et de ce fait a été conçu dans un but premier de reconnaissance des formes. Cependant, il peut aussi être utilisé pour faire de la classification et pour résoudre des opérations logiques simples (telle "ET" ou "OU"). Sa principale limite est qu'il ne peut résoudre que des problèmes linéairement séparables. Il suit généralement un apprentissage supervisé selon la règle de correction de l'erreur (ou selon la règle de Hebb).

✓ **Les perceptrons multicouches :**

C'est une extension du précédent, avec une ou plusieurs couches cachées entre l'entrée et la sortie. Chaque neurone dans une couche est connecté à tous les neurones de la couche précédente et de la couche suivante (excepté pour les couches d'entrée et de sortie) et il n'y a pas de connexions entre les cellules d'une même couche. Les fonctions d'activation utilisées dans ce type de réseaux sont principalement les fonctions à seuil ou sigmoïdes. Il peut résoudre des problèmes non-linéairement séparables et des problèmes logiques plus compliqués, et notamment le fameux problème du XOR. Il suit aussi un apprentissage supervisé selon la règle de correction de l'erreur.

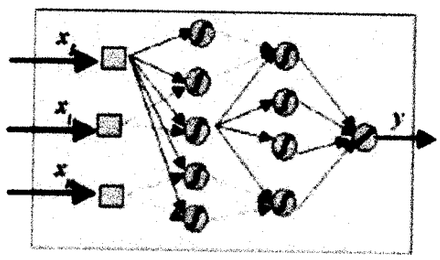


Figure II.11 : Architecture du perceptron multicouche

II.1.5.1.2 Les réseaux à fonction radiale :

Ce sont les réseaux que l'on nomme aussi RBF ("Radial Basic Functions"). L'architecture est la même que pour les PMC cependant, les fonctions de base utilisées ici sont des fonctions Gaussiennes. Les RBF seront donc employés dans les mêmes types de problèmes que les PMC à savoir, en classification et en approximation de fonctions,

particulièrement. L'apprentissage le plus utilisé pour les RBF est le mode hybride et les règles sont soit, la règle de correction de l'erreur soit, la règle d'apprentissage par compétition. (Voir **Annexe A**)

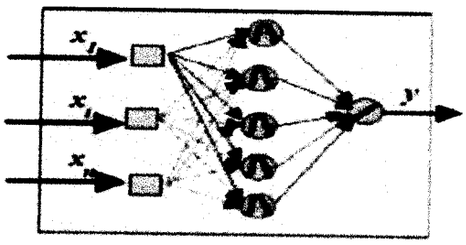


Figure II.12 : Architecture du réseau à fonction radiale

II.1.5.2 LES RESEAUX "FEED-BACK" :

Appelés aussi "réseaux récurrents", ce sont des réseaux dans lesquels il y à retour en arrière de l'information.

- Les réseaux de Hopfield
- Les cartes auto-organisatrices de Kohonen
- Les ART : les réseaux ART ("Adaptative Resonnance Theorie") sont des réseaux apprentissage par compétition.

II.1.6 Apprentissage des réseaux de neurones :

Pour un RNA, l'apprentissage peut être considéré comme le problème de la mise à jour des poids des connexions au sein du réseau, afin de réussir la tâche qui lui est demandée. L'apprentissage est la caractéristique principale des RNA et il peut se faire de différentes manières et selon différentes règles.

II.1.6.1 Types d'apprentissage :

II.1.6.1.1 Le mode supervisé :

Dans ce type d'apprentissage, le réseau s'adapte par comparaison entre le résultat qu'il a calculé, en fonction des entrées fournies, et la réponse attendue en sortie. Ainsi, le réseau va se modifier jusqu'a ce qu'il trouve la bonne sortie, c'est-à-dire celle attendue, correspondant à une entrée donnée. Cet apprentissage est appliqué généralement pour les réseaux non bouclé.

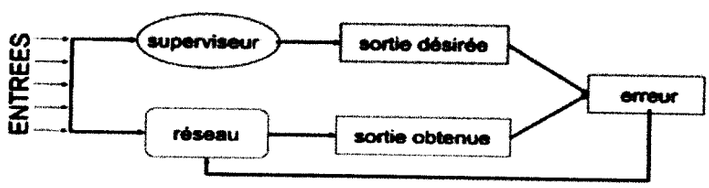


Figure II.13 : Apprentissage supervisé

II.1.6.1.2 Le mode non-supervisé (ou auto-organisationnel) :

L'apprentissage non supervisé est une technique différente ou ne détermine pas de variable de sortie. Le réseau va de lui-même catégoriser les variables d'entrée.

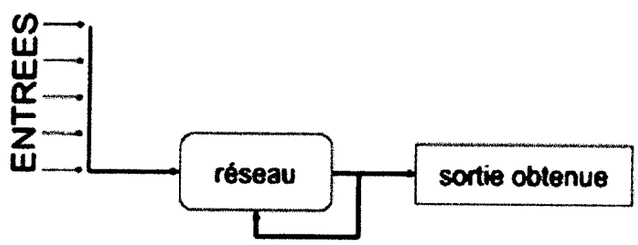


Figure II.14 : Apprentissage non supervisé

Dans ce cas, l'apprentissage est appliqué pour les réseaux bouclé, basé sur des probabilités. Le réseau va se modifier en fonction des régularités statistiques de l'entrée et établir des catégories, en attribuant et en optimisant une valeur de qualité, aux catégories reconnues. On ne sait pas à priori si la sortie obtenue est valable ou non. Les entrées sont projetées sur l'espace de réseau.

II.1.6.1.3 Apprentissage par renforcement :

Le renforcement est en fait une sorte d'apprentissage supervisé et certains auteurs le classe d'ailleurs, dans la catégorie des modes supervisés. Dans cette approche le réseau doit apprendre la corrélation entrée/sortie via une estimation de son erreur, c'est-à-dire du rapport échec/succès. Le réseau va donc tendre à maximiser un index de performance qui lui est fourni, appelé signal de renforcement. Le système étant capable ici, de savoir si la réponse qu'il fournit est correcte ou non, mais il ne connait pas la bonne réponse.

II.1.6.1.4 Le mode hybride:

Le mode hybride reprend en fait les deux autres approches, puisque une partie des poids va être déterminée par apprentissage supervisé et l'autre partie par apprentissage non-supervisé.

[BENDI 2008]

II.1.6.2 Règles d'apprentissage :

- **Règle de correction d'erreurs :**

Cette règle s'inscrit dans le paradigme d'apprentissage supervisé, c'est -à-dire dans le cas où l'on fournit au réseau une entrée et la sortie correspondante. Si on considère y comme étant la sortie calculée par le réseau, et d la sortie désirée, le principe de cette règle est d'utiliser l'erreur (d-y), afin de modifier les connexions et de diminuer ainsi l'erreur globale du système. Le réseau va donc s'adapter jusqu'à ce que y soit égal à d. Ce Principe est notamment utilisé dans le modèle du perceptron simple.

➢ **Apprentissage de Boltzmann :**

Les réseaux de Boltzmann sont des réseaux symétriques récurrents. Ils possèdent deux sous-groupes de cellules, le premier étant relié à l'environnement (cellules dites visibles) et le second ne l'étant pas (cellules dites cachées). Cette règle d'apprentissage est de type stochastique (= qui relève partiellement du hazard) et elle consiste à ajuster les poids des connexions, de telle sorte que l'état des cellules visibles satisfasse une distribution probabiliste souhaitée.

➢ **Apprentissage « local » (Règles de HEBB) :**

Comme je l'ai déjà dit dans l'historique, cette règle, basée sur des données biologiques, modélise le fait que si des neurones, de part et d'autre d'une synapse, sont activés de façon synchrone et répétée, la force de la connexion synaptique va aller croissant. Il est à noté ici que l'apprentissage est localisé, c'est-à-dire que la modification d'un poids synaptique wij ne dépend que de l'activation d'un neurone i et d'un autre neurone j.

➢ **Règle d'apprentissage par compétitions :**

La particularité de cette règle, c'est qu'ici l'apprentissage ne concerne qu'un seul neurone. Le principe de cet apprentissage est de regrouper les données en catégories. Les patrons similaires vont donc être rangés dans une même classe, en se basant sur les corrélations des données, et seront représentés par un seul neurone, on parle de « winner-take-all ». Dans un réseau à compétition simple, chaque neurone de sortie est connecté aux neurones de la couche d'entrée, aux autres cellules de la couche de sortie (connexions inhibitrices) et à elle-même (connexion excitatrice). La sortie va donc dépendre de la compétition entre les connexions inhibitrices et excitatrices. **[RN COUR]**

II.1.7 Applications :

Se trouvant à l'intersection de différents domaines (informatique, électronique, science cognitive, neurobiologie et même philosophie), l'étude des réseaux de neurones est une voie prometteuse de l'Intelligence Artificielle, qui a des applications dans de nombreux domaines :

- **Industrie:** contrôle qualité, diagnostic de panne, corrélations entre les données fournies par différents capteurs, analyse de signature ou d'écriture manuscrite...
- **Finance:** prévision et modélisation du marché (cours de monnaies...), sélection d'investissements, attribution de crédits...
- **Télécommunications et informatique:** analyse du signal, élimination du bruit, reconnaissance de formes (bruits, images, paroles), compression de données...
- **Environnement:** évaluation des risques, analyse chimique, prévisions et modélisation météorologiques, gestion des ressources...
- **Aérospatial et automobile** : pilotage automatique, simulation du vol, système de guidage automatique…
- **Défense** : guidage de missile, suivi de cible, reconnaissance du visage, radar, sonar...
- **Secteur médical** : diagnostic médicale, analyse des l'EEC et l'ECG.

II.1.8 Limites :

- Les réseaux de neurones artificiels ont besoin de cas réels servant d'exemples pour leur apprentissage (on appelle cela la *base d'apprentissage*). Ces cas doivent être d'autant plus nombreux que le problème est complexe et que sa topologie est peu structurée. Par exemple, on peut optimiser un système neuronal de lecture de caractères en utilisant le découpage manuel d'un grand nombre de mots écrits à la main par de nombreuses personnes. Chaque caractère peut alors être présenté sous la forme d'une image brute, disposant d'une topologie spatiale à deux dimensions, ou d'une suite de segments presque tous liés. La topologie retenue, la complexité du phénomène modélisé, et le nombre d'exemples doivent être en rapport. Sur un plan pratique, cela n'est pas toujours facile car les exemples peuvent être soit en quantité absolument limitée ou trop onéreux à collecter en nombre suffisant.
- Il y a des problèmes qui se traitent bien avec les réseaux de neurones, en particulier ceux de *classification* en *domaines convexes* (c'est-à-dire tels que si des points A et B font partie du domaine, alors tout le segment AB en fait partie aussi). Des problèmes comme "*Le nombre d'entrées à 1 (ou à zéro) est-il pair ou impair ?*" se résolvent en revanche très mal : pour affirmer de telles choses sur 2 puissance N points, si on se contente d'une approche naïve mais homogène, il faut précisément N-1 couches de neurones intermédiaires, ce qui nuit à la généralité du procédé.
- Un exemple caricatural, mais significatif est le suivant : disposant en entrée du seul poids d'une personne, le réseau doit déterminer si cette personne est une femme ou un bien un homme. Les femmes étant statistiquement un peu plus légères que les hommes, le réseau fera toujours un peu mieux qu'un simple tirage au hasard : cet exemple dépouillé indique la simplicité et les limitations de ces modèles mais il montre également comment l'étendre : L'information "port d'une jupe", si on l'ajoute, aurait clairement un coefficient synaptique plus grand que la simple information de poids. **[RN WIKI]**

II.1.9 Conclusions :

Les réseaux de neurones sont depuis quelque temps un point de focalisation des médias, du public et des scientifiques. Les travaux menés dans le domaine des sciences de la cognition artificielle ont été marquées par quelques apports non négligeables mais surtout par beaucoup d'optimisme. Les années qui viennent concrétiseront cet optimisme ou bien relègueront cette technique parmi les nombreuses " recettes " informatiques.

II.2 Les algorithmes génétiques :

II.2.1 Introduction :

L'Optimisation est l'une des branches les plus importantes des mathématiques appliquées modernes, et de nombreuses recherches, à la fois pratiques et théoriques, lui sont consacrées. Si on met de côté les problèmes d'optimisation discrète ou multicritère, alors la théorie de l'optimisation peut être séparée en deux grandes branches : l'optimisation locale et l'optimisation globale. Si on peut considérer que la première est presque « entièrement connue », la seconde est encore partiellement méconnue et les recherches y sont à leur apogée, comme le confirment les nombreuses parutions récentes. La tâche principale de l'optimisation globale est la recherche de la solution qui minimisera un critère de coût donné, appelée « optimum global ». L'optimisation globale vise donc à rechercher non seulement un minimum local, mais surtout le plus petit de ces minima locaux.

Il existe deux grandes approches à l'optimisation globale. L'une est dite déterministe : les algorithmes de recherche utilisent toujours le même cheminement pour arriver à la solution, et on peut donc « déterminer » à l'avance les étapes de la recherche. L'autre est aléatoire : pour des conditions initiales données, l'algorithme ne suivra pas le même cheminement pour aller vers la solution trouvée, et peut même proposer différentes solutions. C'est vers cette seconde branche, la recherche globale aléatoire, que vont s'orienter nos travaux, et plus particulièrement vers un type bien précis d'algorithme de recherche aléatoire, les algorithmes génétiques.

II.2.2 Généralités sur les algorithmes génétiques :

Dans la nature, les êtres vivants croissent et interagissent les uns avec les autres. Chaque individu est caractérisé par un génotype indépendant de l'environnement où il vit. Les opérateurs génétiques fonctionnent au niveau génotypique tandis que le mécanisme de sélection opère au niveau phénotypique (le phénotype d'un individu est l'ensemble des traits caractéristiques d'un individu, alors que le génotype est le codage de ces traits en gènes). Les algorithmes génétiques sont à la base des algorithmes d'optimisation stochastiques, mais peuvent également servir pour l'apprentissage automatique, par exemple. Les premiers travaux dans ce domaine ont commencé dans les années cinquante, lorsque plusieurs biologistes américains ont simulé des structures biologiques sur ordinateur. Puis, entre 1960 et 1970, John Holland, sur la base des travaux précédents, développe les principes fondamentaux des algorithmes génétiques dans le cadre de l'optimisation mathématique. Malheureusement, les ordinateurs de l'époque n'étaient pas assez puissants pour envisager l'utilisation des algorithmes génétiques sur des problèmes réels de grande taille. La parution en 1989 de l'ouvrage de référence écrit par D.E. Goldberg qui décrit l'utilisation de ces algorithmes dans le cadre de résolution de problèmes concrets a permis de mieux faire connaître ces derniers dans la communauté scientifique et a marqué le début d'un nouvel intérêt pour cette technique d'optimisation,

qui reste néanmoins très récente. Parallèlement, des techniques proches ont été élaborées, dont on peut notamment citer la programmation génétique ou les stratégies évolutionnistes. Dans le même temps, la théorie mathématique associée aux algorithmes génétiques s'est développée mais reste encore bien limitée face à la complexité théorique induite par ces algorithmes.

Comme nous l'avons mentionné précédemment, les algorithmes génétiques s'attachent à simuler le processus de sélection naturelle dans un environnement hostile lié au problème à résoudre, en s'inspirant des théories de l'évolution proposées par Charles Darwin :

1. Dans chaque environnement, seules les espèces les mieux adaptées perdurent au cours du temps, les autres étant condamnées à disparaître.
2. Au sein de chaque espèce, le renouvellement des populations est essentiellement dû aux meilleurs individus de l'espèce.

On parlera ainsi d'individu dans une population et bien souvent l'individu sera résumé par un seul chromosome (individu haploïde). Les chromosomes sont eux même constitués de gènes qui contiennent les caractères héréditaires de l'individu. On retrouvera aussi les principes fondamentaux de l'évolution naturelle, à savoir les principes de sélection, de croisement, de mutation, etc.

Dans le cadre de l'optimisation, chaque individu représente un point de l'espace d'état auquel on associe la valeur du critère à optimiser. On génère ensuite une population d'individus aléatoirement pour laquelle l'algorithme génétique s'attache à sélectionner les meilleurs individus tout en assurant une exploration efficace de l'espace d'état. Les algorithmes génétiques diffèrent des algorithmes classiques d'optimisation et de recherche essentiellement en quatre points fondamentaux :

1. Les algorithmes génétiques utilisent un codage des éléments de l'espace de recherche et non pas les éléments eux même.
2. Les algorithmes génétiques recherchent une solution à partir d'une population de points et non pas à partir d'un seul point.
3. Les algorithmes génétiques n'imposent aucune régularité sur la fonction étudiée (continuité, dérivabilité, convexité...). C'est un des gros atouts des algorithmes génétiques.
4. Les algorithmes génétiques ne sont pas déterministes, ils utilisent des règles de transition probabilistes.

La robustesse est une des caractéristiques principales des algorithmes génétiques : ils permettent de fournir une ou plusieurs solutions de « bonne » qualité (pas nécessairement optimales, mais suffisantes en pratique) à des problèmes très variés, en sollicitant un investissement (temps et puissance de calcul) assez faible. En effet, l'heuristique de l'évolution est en quelque sorte « universelle », et très peu d'informations suffisent pour résoudre un problème quelconque.

C'est ainsi qu'ils ont donné de bons résultats dans le domaine médical. Ils sont également efficaces sur des problèmes pour lesquels il n'existe pas encore d'algorithme de résolution ou dont la taille est rédhibitoire pour les méthodes classiques. **[AGGN]**

II.2.3 Principes généraux des algorithmes génétiques :

Un algorithme génétique recherche le ou les extrema d'une fonction définie sur un espace de données. Pour l'utiliser, on doit disposer des cinq éléments suivants :

1. Un principe de codage de l'élément de population. Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. La qualité du codage des données conditionne le succès des algorithmes génétiques. Le codage binaires ont été très utilises à l'origine. Les codages réels sont désormais largement utilises, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.
2. Un mécanisme de génération de la population initiale. Ce mécanisme doit être capable de produire une population d'individus non homogène qui servira de base pour les générations futures. Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global. Dans le cas ou l'on ne connaît rien du problème à résoudre, il est essentiel que la population initiale soit repartie sur tout le domaine de recherche.
3. Une fonction à optimiser. Celle-ci retourne une valeur de $\Re^+$pelée *fitness* ou fonction d'évaluation de l'individu.
4. Des operateurs permettant de diversifier la population au cours des générations et d'explorer l'espace d'état. L'operateur de croisement recompose les gènes d'individus existant dans la population, l'operateur de mutation a pour but de garantir l'exploration de l'espace d''etats.
5. Des paramètres de dimensionnement : taille de la population, nombre total de générations ou critère d'arrêt, probabilités d'application des operateurs de croisement et de mutation.

Le principe général du fonctionnement d'un algorithme génétique est représenté sur la **Figure II.15**:

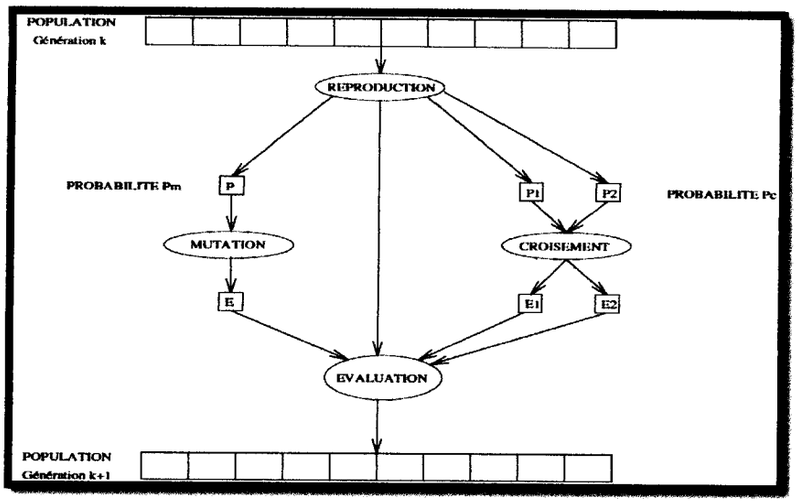


Figure II.15 : Principe général des algorithmes génétiques

On commence par générer une population d'individus de façon aléatoire. Pour passer d'une génération k à la génération k+1, les trois opérations suivantes sont répétées pour tous les éléments de la population k. Des couples de parents P_1 et P_2 sont sélectionnés en fonction de leurs adaptations. L'opérateur de croisement leur est appliqué avec une probabilité P_c (généralement autour de 0.6) et génère des couples d'enfants C_1 et C_2. D'autres éléments P sont sélectionnés en fonction de leur adaptation. L'opérateur de mutation leur est appliqué avec la probabilité P_m (P_m est généralement très inférieur à P_c) et génère des individus mutés P'. Le niveau d'adaptation des enfants ($C1$, $C2$) et des individus mutés P' sont ensuite évalués avant insertion dans la nouvelle population. Différents critères d'arrêt de l'algorithme peuvent être choisis :

- Le nombre de générations que l'on souhaite exécuter peut être fixé à priori. C'est ce que l'on est tenté de faire lorsque l'on doit trouver une solution dans un temps limité.
- L'algorithme peut être arrêté lorsque la population n'évolue plus ou plus suffisamment rapidement. **[LHABIT]**

II.2.3.1 Cycle de base et présentation formelle de l'AG :

On peut présenter de manière synthétique le fonctionnement de l'algorithme génétique par le cycle suivant :

initialiser la population (générer aléatoirement une population de *N* chromosomes *x*)
calculer le degré d'adaptation *f(x)* de chaque individu
Tant que non fini ou non convergence
    reproduction des parents
        sélectionner 2 individus à la fois
        appliquer les opérateurs génétiques
    calculer le degré d'adaptation *f(x)* de chaque enfant
    sélectionner les survivants parmi les parents et les enfants
fin Tant que
conclure

Figure II.16 : Algorithme des AGs

Le critère de convergence peut être de nature diverse, par exemple :

- Un taux minimum qu'on désire atteindre d'adaptation de la population au problème,
- Un certain temps de calcul à ne pas dépasser,
- Une combinaison de ces deux points.

II.2.3.2 Schéma récapitulatif :

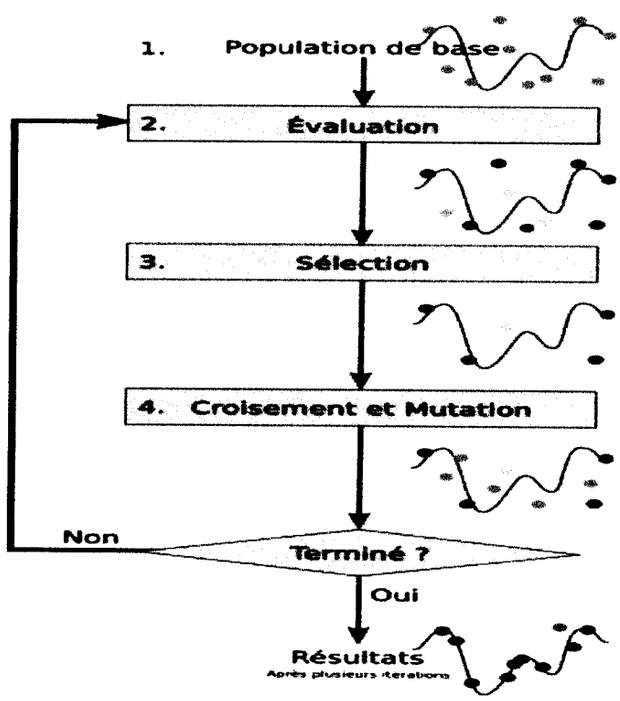


Figure II.17 : Schéma simple d'un AG [WIKIAG]

1. **Population de base générée aléatoirement :**

n chaînes de caractères ou de bits.

1 chaîne correspond à 1 chromosome.

2. **Évaluation :**

À chaque chaîne, une note correspondant à son adaptation au problème.

3. **Sélection :**

Tirage au sort de n/2 couples de chaînes sur une roue biaisée. Chaque chaîne a une probabilité d'être tirée proportionnelle à son adaptation au problème. Optimisation possible : si l'individu le plus adapté n'a pas été sélectionné, il est copié d'office dans la génération intermédiaire à la place d'un individu choisi aléatoirement.

4. **Croisement et mutation :**

Chaque couple donne 2 chaînes filles.

- **Enjambement**. Probabilité : 70%. Emplacement du l'enjambement choisi aléatoirement. Exemple :
  Chaînes parents : A : 00110100 ; B : 01010010
  Chaînes filles : A' : 00010010 ; B' : 01110100
  Croisement en 2 points plus efficace.

- **Mutations des chaînes filles**. Probabilité : de 0,1 à 1%.
  Inversion d'un bit au hasard ou remplacement au hasard d'un caractère par un autre.
  Probabilité fixe ou évolutive (auto-adaptation).
  On peut prendre probabilité = 1/nombre de bits. **[WIKIAG]**

II.2.4 Les différentes étapes des algorithmes génétiques :

II.2.4.1 Codage des données (des individus) :

Dans la nature, les structures géniques sont codées en base 4, dont les "chiffres" sont les quatre bases azotées : l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G). Dans le cadre des algorithmes génétiques, ce type de codage est bien difficile à utiliser et n'est donc pas retenu. Le principal problème qui se pose est celui concernant la conservation de la topologie du problème par le codage. **[AGIA]**

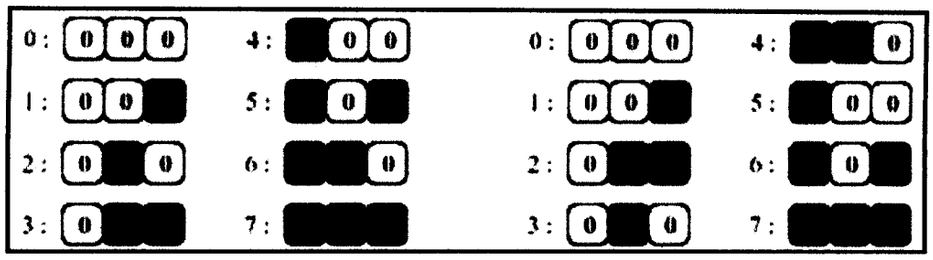


Figure II.18 : Codage binaire classique et codage de Gray

Historiquement le codage utilisé par les algorithmes génétiques était représenté sous forme de chaînes de bits contenant toute l'information nécessaire à la description d'un point dans l'espace d'état. Ce type de codage a pour intérêt de permettre de créer des opérateurs de croisement et de mutation simples (par inversion de bits par exemple). C'est également en utilisant ce type de codage que les premiers résultats de convergence théorique ont été obtenus.

Cependant, ce type de codage n'est pas toujours bon comme le montrent les deux exemples suivants :

- Deux éléments voisins en terme de distance de Hemming (représente le nombre de bits dont diffèrent deux nombres binaires) ne codent pas nécessairement deux éléments proches dans l'espace de recherche : par exemple, dans le cas présenté sur la figure ci-dessus, les individus 1 et 2 sont proches mais ne le sont plus après transposition en binaire (« 001 » et « 010 », donc une distance de Hemming de 2), alors que « 1 » et « 3 » deviennent proches après codage binaire (« 001 » et « 011 »). On peut cependant remédier à ceci en utilisant le codage réfléchi (ou codage de Gray) qui conserve une distance de Hemming de « 1 » entre deux individus consécutifs quelconques, donc la topologie du problème.
- pour des problèmes d'optimisation dans des espaces de grande dimension, le codage binaire peut rapidement devenir mauvais. En effet, l'ordre des variables a une importance dans la structure du chromosome binaire, alors qu'il n'en a pas forcément dans la structure du problème. Enfin, la structure binaire empêche l'utilisateur d'accéder à une valeur particulière.

Les algorithmes génétiques utilisant des vecteurs réels évitent ce problème en conservant les variables du problème dans le codage de l'élément de population sans passer par le codage binaire intermédiaire. La structure du problème est conservée dans le codage.

II.2.4.2 Génération aléatoire de la population initiale :

Le choix de la population initiale d'individus conditionne fortement la rapidité de l'algorithme. Si la position de l'optimum dans l'espace d'état est totalement inconnue, il est naturel de générer aléatoirement des individus en faisant des tirages uniformes dans chacun des domaines associés aux composantes de l'espace d'état en veillant à ce que les individus produits respectent les contraintes .

Si par contre, des informations à priori sur le problème sont disponibles, il parait bien évidemment naturel de générer les individus dans un sous-domaine particulier afin d'accélérer la convergence.

Dans l'hypothèse où la gestion des contraintes ne peut se faire directement, les contraintes sont généralement incluses dans le critère à optimiser sous forme de pénalités. Il est clair qu'il vaut mieux, lorsque c'est possible ne générer que des éléments de population respectant les contraintes. **[LHABIT]**

II.2.4.3 Gestion des contraintes :

Un élément de population qui viole une contrainte se verra attribuer une mauvaise fitness et aura une probabilité forte d'être éliminé par le processus de sélection.

Il peut cependant être intéressant de conserver, tout en les pénalisant, les éléments non admissibles car ils peuvent permettre de générer des éléments admissibles de bonne qualité. Pour de nombreux problèmes, l'optimum est atteint lorsque l'une au moins des contraintes de séparation est saturée, c'est à dire sur la frontière de l'espace admissible.

Gérer les contraintes en pénalisant la fonction fitness est difficile, un « dosage » s'impose pour ne pas favoriser la recherche de solutions admissibles au détriment de la recherche de l'optimum ou inversement.

Disposant d'une population d'individus non homogène, la diversité de la population doit être entretenue au cours des générations afin de parcourir le plus largement possible l'espace d'état. C'est le rôle des opérateurs de croisement et de mutation. **[BENDI 2008]**

II.2.4.4 Les opérateurs :

Les opérations successives utilisées dans les algorithmes génétiques sont illustrées sur la figure ci-dessus : Sélection, Mutation, Croisement (cross-over)

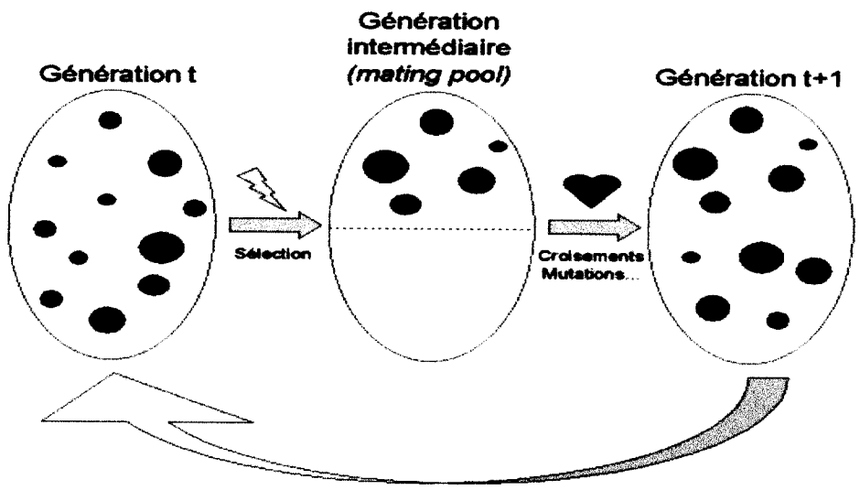


Figure II.19 : Les opérateurs utilisés dans les AG [AGIA]

II.2.4.4.1 Opérateur de sélection : [LHABIT]

A l'inverse d'autres techniques d'optimisation, les algorithmes génétiques n'ont pas besoin de connaître la dérivée de la fonction objectif, ce qui rend leur domaine d'application plus vaste.

La sélection, comme son nom l'indique, permet d'identifier statistiquement les meilleurs individus d'une population et d'éliminer partiellement les mauvais. Néanmoins, comme dans la Nature, il ne faut pas confondre sélection et élitisme : ce n'est pas parce qu'un individu est bon qu'il survivra nécessairement (les aléas de la vie) et de même ce n'est pas parce qu'il est mauvais qu'il doit disparaître (la chance aide également à survivre). En effet, bien souvent, une espèce « bien adaptée » peut descendre d'un individu décrété « mauvais ». On distingue plusieurs méthodes de sélection :

a. La roulette :

La « *roulette Wheel selection* » (ou **sélection par roulette de casino**) : elle consiste à associer à chaque individu un segment dont la longueur est proportionnelle à sa fitness. Ces segments sont ensuite concaténés sur un axe gradué que l'on normalise entre 0 et 1. On tire alors un nombre aléatoire de distribution uniforme entre 0 et 1, puis on regarde quel est le segment sélectionné, et on reproduit l'individu correspondant. Avec cette technique, les bons individus seront plus souvent sélectionnés que les mauvais, et un même individu pourra avec cette méthode être sélectionné plusieurs fois. Néanmoins, sur des populations de petite taille, il est difficile d'obtenir exactement l'espérance mathématique de sélection à cause du faible nombre de tirages (le cas idéal d'application de cette méthode est bien évidemment celui où la population est de taille infinie). On aura donc un biais de sélection plus ou moins fort suivant la dimension de la population.

Individus	*fitness*	*fitness relative*
1	15	0.1
2	30	0.2
3	45	0.3
4	22.5	0.15
5	7.5	0.05
6	30	0.2

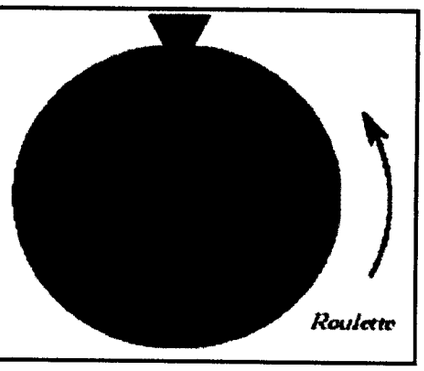


Figure II.20 : Exemple d'application de la roulette Wheel selection

b. Stochastique remainder :

La « *stochastic remainder without replacement selection* » (ou reste stochastique sans choix de remplacement) : elle évite ce genre de problème, car une partie de la population

est sélectionnée de manière purement déterministe. On associe à chaque individu le rapport ri de sa fitness sur la moyenne des fitness puis on prend sa partie entière E(ri) qui indique le nombre de fois à reproduire l'individu i. On assure ainsi un nombre exact de représentants pour la génération suivante, ce qui élimine le biais. Cependant, les individus faibles (fitness inférieure à la fitness moyenne) sont invariablement éliminés avec cette méthode, ce qui est mauvais, car ceux-ci occupent des positions dans l'espace d'état qui associées avec d'autres peuvent nous rapprocher du sous-domaine contenant l'optimum. On associe donc à la sélection déterministe un principe de sélection aléatoire basée sur une roulette Wheel selection exécutée sur tous les individus affectés de nouvelles finesses ri-E (ri).

Individus	fitness	r_i	$E(r_i)$	$r_i - E(r_i)$	p_i
1	15	0.6	0	0.6	1/5
2	30	1.2	1	0.2	1/15
3	45	1.8	1	0.8	4/15
4	22.5	0.9	0	0.9	3/10
5	7.5	0.3	0	0.3	1/10
6	30	1.2	1	0.2	1/15

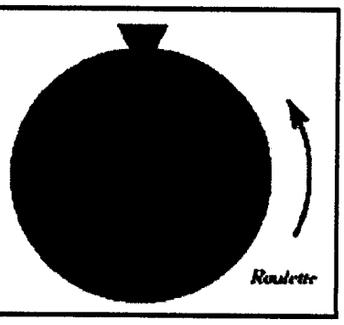


Figure II.21 : Application de la stochastic remainder without replacement selection à l'exemple précédent

On ajoute également fréquemment un principe d'élitisme dans le processus de sélection destiné à conserver systématiquement le ou les meilleurs individus de la population courante dans la génération suivante, sans lui faire subir de croisement ou de mutation qui pourraient le détruire. Ce principe confère aussi à l'algorithme génétique la propriété de croissance monotone de la fitness du meilleur individu au cours des générations.

Les résultats issus de la théorie des schémas montrent que le principe de la *roulette Wheel sélection* offre le meilleur compromis entre exploration de l'espace de recherche et exploitation des informations obtenues. Cependant, les hypothèses nécessaires à ce résultat sont rarement satisfaites en pratique et les autres principes de sélection sont souvent plus efficaces.

c. Sélection par tournoi :

A chaque fois qu'il faut sélectionner un individu, la « sélection par tournoi » consiste à tirer aléatoirement (k) individus de la population, sans tenir compte de la valeur de leur fonction d'adaptation, et de choisir le meilleur individu parmi les k individus. Le nombre d'individus sélectionnés a une influence sur la pression de sélection, lorsque $k = 2$, la sélection est dite par « tournoi binaire ».

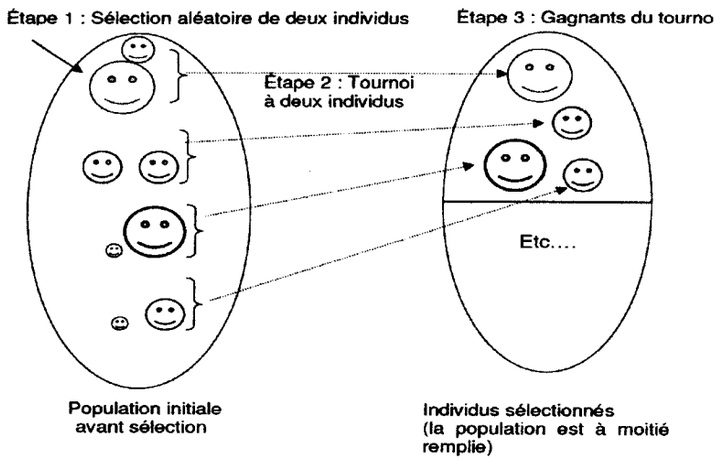


Figure II.22 : Représentation d'une sélection par tournoi d'individus pour un critère de maximisation. Chaque individu représente une solution possible [TROAG]

d. Sélection uniforme :

C'est une technique très simple qui consiste à sélectionner un individu Ci aléatoirement de la population P. La probabilité pi pour qu'un individu soit sélectionné est définie par :

$$Pi = (1/taille\ pop)$$

e. Stochastique uniforme :

Dans cette méthode on dispose une ligne dont lequel chaque parents correspond à une longueur proportionnel à sa valeur. L'algorithme se déplace la long de la ligne dans des pas de taille égaux a chaque pas l'algorithme prend un parent.

II.2.4.4.2 Opérateur de Croisement :

Le croisement a pour but d'enrichir la diversité de la population en manipulant la structure des chromosomes. Classiquement, les croisements sont envisagés avec deux parents et génèrent deux enfants. Il consiste à échanger les gènes des parents afin de donner des enfants qui portent des propriétés combinées. Bien qu'il soit aléatoire, cet échange d'informations offre aux algorithmes génétiques une part de leur puissance : quelque fois, de " bons " gènes d'un parent viennent remplacer les " mauvais " gènes d'un autre et créent des fils mieux adaptés aux parents.

Il existe différentes techniques de croisement. Chaque des techniques s'applique sur des chromosomes dont la représentation est soit binaire ou réelle. Nous citerons quelque technique :

a. Croisement classique et à 2 point :

Initialement, le croisement utilisé avec les chaînes de bits était le croisement à découpages de chromosomes, ou *slicing cross over*. Pour effectuer ce type de croisement sur des chromosomes constitués de L gènes, on tire aléatoirement une position inter-gènes dans chacun des parents. On échange ensuite les deux sous-chaînes de chacun des chromosomes, ce qui produit deux enfants C1 et C2. Ce mécanisme présente l'inconvénient de privilégier les extrémités des individus. Et selon le codage choisi, il peut générer des fils plus ou moins proches de leurs parents. Pour éviter ce problème, on peut étendre ce principe en découpant le chromosome non pas en 2 sous-chaînes mais en 3, 4, etc. On parle alors de *k-point slicing cross over* (figures ci-dessous).

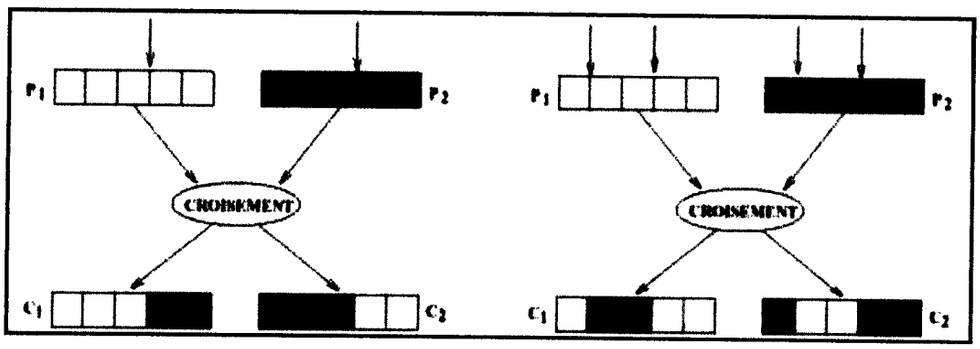


Figure II.23 : a. Slicing cross over classique b. Slicing cross over à 2 points

b. Croisement barycentrique :

Ce type de croisement à découpage de chromosomes est très efficace pour les problèmes discrets. Pour les problèmes continus, un croisement « barycentrique » est souvent utilisé : deux gènes $P_1(i)$ et $P_2(i)$ sont sélectionnés dans chacun des parents à la même position i. Ils définissent deux nouveaux gènes $C_1(i)$ et $C_2(i)$ par combinaison linéaire :

$$\begin{cases} C_1(i) = \alpha P_1(i) + (1-\alpha) P_2(i) \\ C_2(i) = (1-\alpha) P_1(i) + \alpha P_2(i) \end{cases}$$

Où α est un coefficient de pondération aléatoire adapté au domaine d'extension des gènes (il n'est pas nécessairement compris entre 0 et 1, il peut par exemple prendre des valeurs dans l'intervalle [-0.5, 1.5] ce qui permet de générer des points entre, ou à l'extérieur des deux gènes considérés).

Dans le cas particulier d'un chromosome matriciel constitué par la concaténation de vecteurs, on peut étendre ce principe de croisement aux vecteurs constituant les gènes :

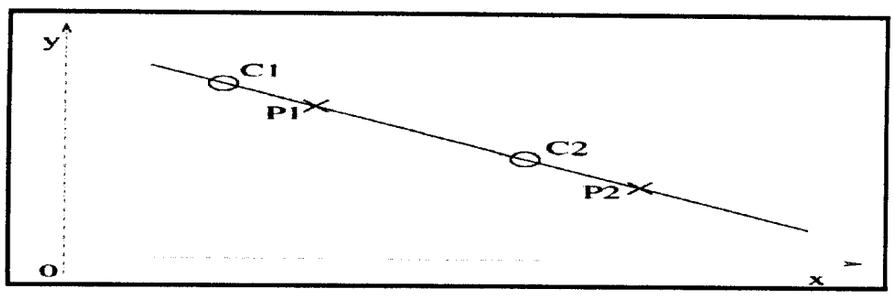


Figure II.24 : Croisement barycentrique

On peut imaginer et tester des opérateurs de croisement plus ou moins complexes sur un problème donné mais l'efficacité de ce dernier est souvent liée intrinsèquement au problème.

c. Croisement étendu (inter médiate) :

Le gène hi du chromosome *H* généré par cet opérateur est défini comme suit :

$$Hi = ci^1 + \alpha_i (ci^2 - ci^1)$$

Tels que la valeur de a est choisie aléatoirement ayant une distribution uniforme de l'intervalle [-d, 1+d]. Dans ce type de croisement, Muhlenbein a déterminé la valeur de d à 0.Dans le cas contraire ce type de croisement est appelé le croisement intermédiaire étendu. Le bon choix de la valeur de d est 0.25 et la valeur de a varie pour chaque gène. L'opérateur de Muhlenbein ne s'applique que sur des gènes de type réel. Autrement dit la formule précédente s'écrit : $hi = ci^1 + \alpha_i (ci^2 - ci^1)$

Nous remarquons que les valeurs des gènes des chromosomes générés n'appartiennent plus à l'intervalle [ai, bi]. Ce qui fait que le domaine d'exploration est assez vaste. **[sekkal2009]**

d. Croisement heuristique :

La même technique que le croisement étendu seulement $hi = ci^2 + \alpha_i (ci^2 - ci^1)$

II.2.4.4.3 Opérateur de mutation :

L'opérateur de mutation apporte aux algorithmes génétiques la propriété d'ergodicité de parcours de l'espace. Cette propriété indique que l'algorithme génétique sera susceptible d'atteindre tous les points de l'espace d'état (sans pour autant nécessairement les adresser tous dans le processus de résolution). En toute rigueur, l'algorithme peut converger sans croisement, et certaines variantes fonctionnent de cette manière, et les propriétés de convergence des algorithmes génétiques sont donc fortement dépendantes de cet opérateur.

a. mutation discrète et continue :

Pour des problèmes discrets, l'opérateur de mutation consiste généralement à tirer aléatoirement un gène dans le chromosome et à remplacer ce dernier par une valeur tirée aussi aléatoirement de l'alphabet propre au gène sélectionné. Dans le cas du codage binaire, la méthode classique consiste, après avoir déterminé le locus à muter, d'appliquer un *non* logique à la valeur de l'allèle correspondant. Bien que ce déplacement puisse paraître petit au niveau des chaînes de bits, il peut être assez important dans la topologie initiale (considérons une mutation transformant " 000 " en " 100 " avec une topologie initiale basée sur la valeur décimale de la chaîne de bits). Ceci introduit un risque de ne pas générer la descendance dans le voisinage des parents.

Dans les problèmes continus, on procède un peu de la même manière en tirant aléatoirement un gène dans le chromosome, auquel on ajoute un bruit aléatoire (par exemple un bruit gaussien) en veillant bien évidemment à ce que le gène résultant reste dans le domaine de définition qui lui est propre. L'écart type de ce bruit est délicat et difficile à régler : s'il est trop faible, l'exploration sera ralentie au début et on risque la convergence locale ; s'il est trop grand, les solutions seront modifiées trop brutalement et on ne pourra pas non plus converger localement vers l'optimum.

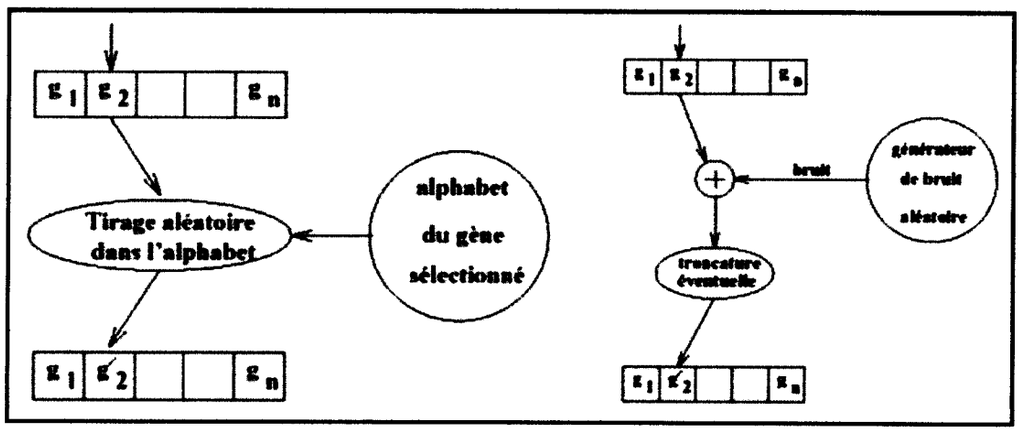


Figure II.25 : a. mutation discrète **b. mutation continue**

Comme dans le cas du croisement, on peut imaginer de concentrer les mutations sur les gènes faibles lorsque le problème présente une fitness construite à l'aide d'un ensemble de sous-fitness associées à chacun des gènes.

b. mutation auto-adaptative :

Il existe également un principe de mutation adaptative, permettant d'optimiser le taux de mutation en codant ce dernier dans la structure du chromosome. Ce second chromosome est géré de la même manière que le premier chromosome codant l'espace d'état, c'est-à-dire lui-même soumis aux opérateurs génétiques (croisement et mutation). Au cours du déroulement de l'algorithme, les gènes et les individus ayant des probabilités de mutation

élevées auront tendance à disparaître à mesure que la population converge vers l'optimum. De même, les gènes ayant des probabilités de mutation trop faibles ne peuvent évoluer favorablement et tendent à être supplantés.

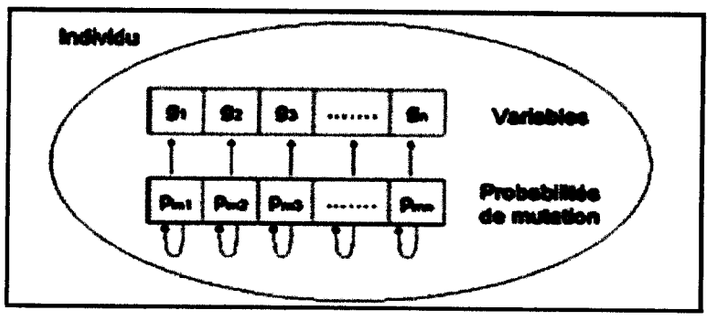


Figure II.26 : Principe de la mutation auto-adaptative

Des opérateurs de mutation très efficaces qui effectuant une optimisation locale sont fréquemment utilisés. Par exemple, Marc Schoenauer a hybridé un algorithme génétique avec un algorithme déterministe de type Newton utilisé en lieu et place de l'opérateur de mutation, appelé *Surrogate Deterministic Mutation*.

c. Mutation gaussien :

La mutation Gaussienne consiste à ajouter un bruit Gaussien aux composantes du vecteur individu concerné, ce qui implique l'ajustement d'un paramètre supplémentaire δ la déviation standard de ce bruit :

$$\forall\, i \in 1, \dots n, xi + N(q, \delta)$$

L'ajustement de Q est relativement complexe(trop petit, il ralentit l'évolution, trop grand, il perturbe la convergence de l'AG), de nombreuses stratégies ont été proposées, consistant à rendre ce paramètre variable au cours de l'évolution, soit en fonction du temps, de la valeur de fitness, dépendant des axes de l'espace de recherche (mutation non isotropes),ou encore auto-adaptatif, comme ci-après. Des études ont aussi été conduites sur l'emploi de bruits non gaussiens. **[sekkal2009]**

II.2.5 Améliorations classiques :

Les processus de sélection présentés sont très sensibles aux écarts de fitness et dans certains cas, un très bon individu risque d'être reproduit trop souvent et peut même provoquer l'élimination complète de ses congénères; on obtient alors une population homogène contenant un seul type d'individu. Ainsi, dans l'exemple de la figure **II.27** le second mode M_2 risque d'être le seul représentant pour la génération suivante et seule la mutation pourra aider à atteindre l'objectif global M_1 au prix de nombreux essais successifs.

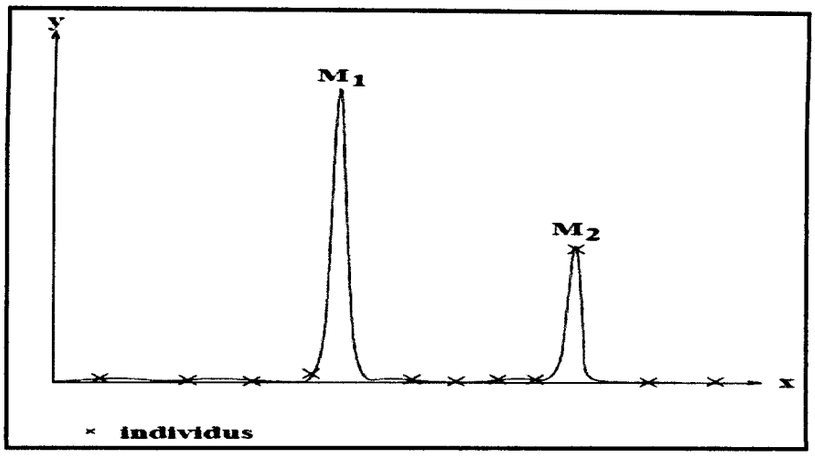


Figure II.27 : Exemple où les sélections classiques risquent de ne reproduire qu'un individu

Pour éviter ce comportement, il existe d'autres modes de sélection (*ranking*) ainsi que des principes (*scaling, sharing*) qui empêchent les individus « forts » d'éliminer complètement les plus « faibles ».

II.2.5.1 Le scaling :

Le *scaling* ou mise à l'échelle, modifie les fitness afin de réduire ou d'amplifier artificiellement les écarts entre les individus. Le processus de sélection n'opère plus sur la fitness réelle mais sur son image après scaling. Parmi les fonctions de scaling, on peut envisager le scaling linéaire et le scaling exponentiel. Soit f_r la fitness avant scaling et f_s la fitness modifiée par le scaling.

A. *Scaling linéaire :*

Dans ce cas la fonction de scaling est définie de la façon suivante :

$$f_s = a f_r + b$$

$$a = \frac{max' - min'}{max - min};\ b = \frac{min'.max - min.max'}{max - min}.$$

En règle générale, le coefficient a est inférieur à un, ce qui permet de réduire les écarts de fitness et donc de favoriser l'exploration de l'espace. Ce scaling est statique par rapport au numéro de génération et pénalise la fin de convergence lorsque l'on désire favoriser les modes dominants.

B. ***Scaling rang (Rankscaling) :*** **[sekkal2009]**

Dans Rankscaling, chaque chromosome se voit associé un rang en fonction de sa position. Le plus mauvais chromosome aura le rang 1, le suivant 2, et ainsi de suite jusqu'au meilleur chromosome qui aura le rang N (pour une population de N chromosomes). Avec cette méthode, tous les chromosomes ont une chance d'être sélectionnés. Cependant, elle conduit à une convergence plus lente vers la bonne solution. Ceci est du au fait que les meilleurs chromosomes ne diffèrent pas énormément des plus mauvais.

C. Scaling proportionnel :

Cette technique fait tourne les résultats autour de sa moyenne fs= 2*moyenne (fr)-fr
Scaling proportionnel à des inconvenant quand le bon individu n'est dans une bonne gamme.

II.2.5.2 Sharing :

L'objectif du sharing est de répartir sur chaque sommet de la fonction à optimiser un nombre d'individus proportionnel à la fitness associée à ce sommet. La figure II.28 présente deux exemples de répartitions de populations dans le cas d'une fonction à cinq sommets : le premier sans sharing, le second avec sharing.

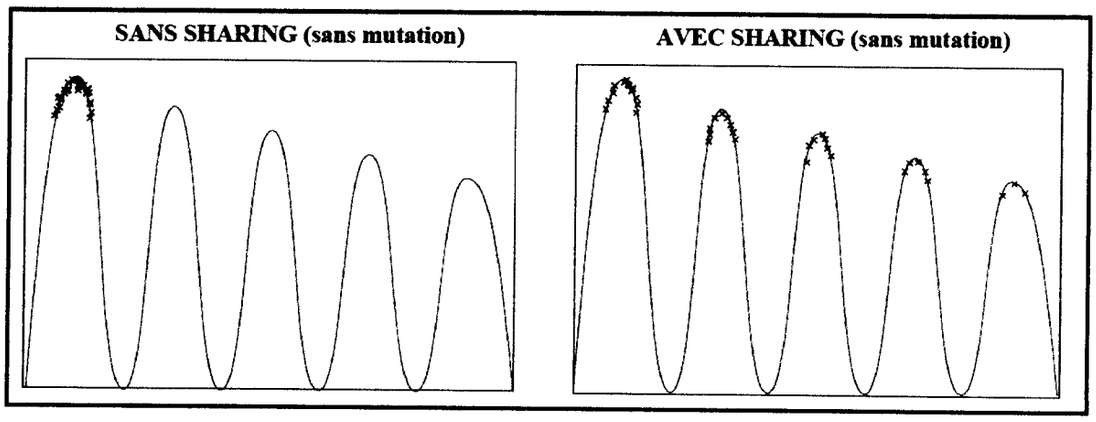


Figure II.28 : Objectif du sharing

De la même façon que le scaling, le sharing consiste à modifier la fitness utilisée par le processus de sélection. Pour éviter le rassemblement des individus autour d'un sommet dominant, le sharing pénalise les fitness en fonction du taux d'agrégation de la population dans le voisinage d'un individu. Il requiert l'introduction d'une notion de distance.

II.2.6 Avantages et inconvénients des AG :

II.2.6.1 Les avantage des AG :

- Les AG opèrent au niveau du codage des paramètres sans se soucier de leur nature, donc ils s'appliquent à de nombreuses classes de problèmes, problèmes qui dépendent éventuellement de plusieurs paramètres de plusieurs paramètres de natures différentes (Booléens, entiers, réels, fonctions…).
- Pour les mêmes raisons, un AG est dans l'idéal totalement indépendant de la nature du problème et de la fonctionnelle à optimiser, car il ne se sert que des valeurs d'adaptation, qui peuvent être très différentes des valeurs de la fonction optimisée, même si elles sont calculées à partir de cette dernière.
- Potentiellement, les AG explorent tout l'espace des points en même temps, ce qui limite les risques de tomber dans des optimums locaux.
- Les AG ne se servent que des valeurs de la fonctionnelle pour optimiser cette dernière, il n'y a pas besoin d'effectuer de couteux et parfois très complexes calculs de dérivées par exemple.

II.2.6.2 Les inconvénients des AG :

- Les AG ne sont encore actuellement pas très en cout (ou vitesse de convergence), vis-à-vis de méthodes d'optimisation plus classiques.
- Le respect des contraintes de domaine par les solutions codées sous forme de chaines de bits pose parfois problèmes. Il faut bien choisir le codage, voire modifier les opérateurs.
- En pratique l'efficacité d'un AG dépend souvent de la nature du problème d'optimisation. selon les cas, le choix des paramètres et des opérateurs sera souvent critique, mais aucune théorie générale ne permet de connaitre avec certitude la bonne paramétrisation. Il faudra faire plusieurs expériences pour s'en approcher. **[AVANT]**

II.2.7 Conclusion :

Les algorithmes génétiques ont pour but de résoudre de tels problèmes par leur approche spécifique, différente des algorithmes d'optimisation les plus courants. Le fait de travailler sur une population implique un parallélisme implicite : ce sont plusieurs solutions qui sont explorées simultanément. De plus, il est possible d'arrêter à tout moment un tel algorithme, il propose toujours une solution, qui n'est pas forcément la meilleure, mais qui n'est pas trop mauvaise non plus.

Enfin les algorithmes génétiques évitent un piège très souvent rencontré dans les algorithmes d'optimisation : ils ne s'arrêtent pas dans les extrema locaux, c'est-à-dire qu'ils essayent constamment de trouver de meilleures solutions, même s'ils semblent les avoir atteintes. **[CONCLUAG]**

II.3 Hybridation algorithme génétique et réseaux de neurones :

II.3.1 Introduction :

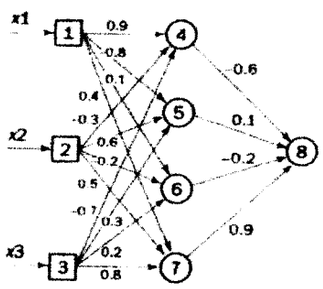
Le concept de système hybride ou de méthode hybride est très large. Ce groupe d'applications inclut toute méthode qui intègre au moins deux approches différentes pour la solution d'un problème donné. Cette hybridation permet de tirer les avantages des deux techniques hybridées. Plusieurs approches peuvent être hybridées :

- ✓ système symbolique-flous,
- ✓ systèmes symbolique-génétique,
- ✓ système Génético-flous,
- ✓ système neuro-flou,
- ✓ système neuro-symboliques,
- ✓ ***système neuro-génétique.***

II.3.2 Utilisation des AG pour une optimisation des poids :

Les parties suivantes présentent le concept de base d'une technique d'optimisation de poids génétiques. Pour une utilisation des algorithmes génétiques, il faut d'abord représenter le domaine de problème comme un chromosome.

Par exemple, nous voulons optimiser les poids d'un perceptron multicouche présenté dans la figure II.29 suivante :



Chromosome:

0.9	-0.3	-0.7	-0.8	0.6	0.3	0.1	-0.2	0.2	0.4	0.5	0.8	-0.6	0.1	-0.2	0.9

Figure II.29 : Présentation d'un système neuro-génétique

Dans **la première étape** on va générer Des poids initiaux dans le réseau choisi aléatoirement dans le petit intervalle **[-1,1]**. Dans ce perceptron, il y a 16 liaisons pondérées entre les neurones .puisqu'un chromosome est un ensemble de gènes, l'ensemble des poids peut être représenté par un chromosome à 16 gènes, où chaque gène

correspond à une liaison simple pondérées dans le réseau .Ce chromosome présente un individu d'une population c.ad une solution proposé à partir d'un ensemble des solutions.

Dans **la deuxième étape** on doit définir une fonction d'évaluation (fitness) pour évaluer la performance des chromosomes. Cette fonction doit estimer la performance d'un réseau neuronal donné. Nous pouvons appliquer ici une fonction assez simple définie par la réciproque de l'erreur quadratique telle que l'algorithme génétique essaye de trouver un ensemble des poids (individu) qui réduisent au minimum la somme d'erreurs quadratique

On peut utiliser aussi comme une fonction le taux de classification non correcte et l'algorithme génétiques essaye de trouvé l'individu qui réduise aux minimum ce taux.

La troisième étape on doit appliquer les deux opérateurs des algorithmes génétiques croissement et mutation. Un opérateur des croissement prend deux chromosomes parentaux et crée un enfant simple avec le matériel génétique des deux parents. Chaque gène dans le chromosome de l'enfant est représenté par vous la transmission des parents aléatoirement choisi. La figure II.30 montre une application de l'opérateur de croisement.

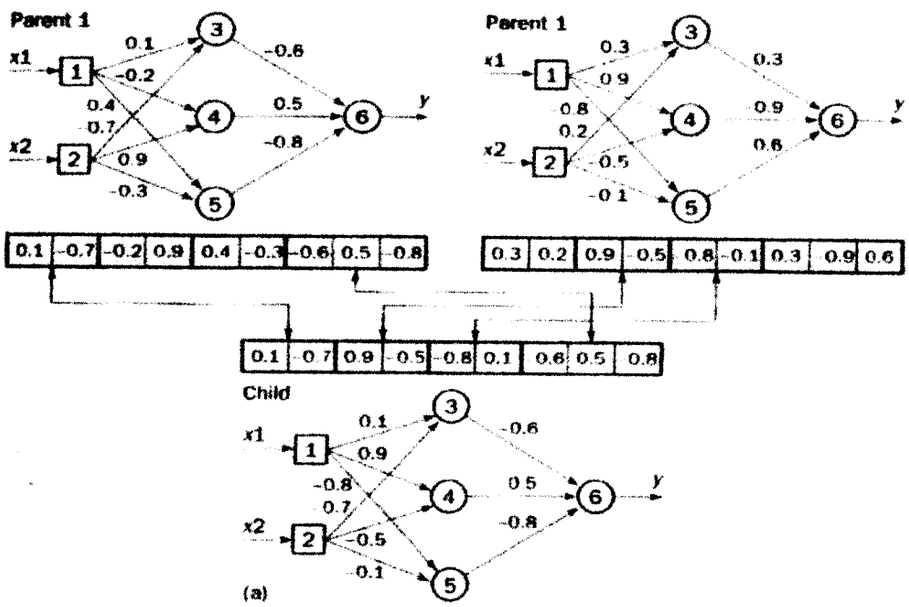


Figure II.30 : Opérateur de croisement dans un système neuro-génétique

Un opérateur de mutation choisit aléatoirement un gène dans un chromosome et ajoute une petite valeur aléatoire à chaque poids dans ce gène. La figure II.31 montre un exemple de mutation.

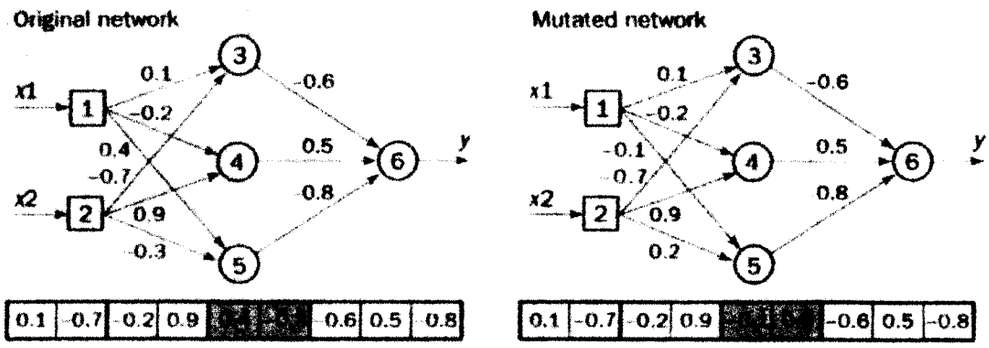


Figure II.31 : Opérateur de mutation dans le système neuro-génétique

Maintenant nous sommes prêts à appliquer l'algorithme génétique. Bien sûr, nous devons toujours définir la taille de population, c'est-à-dire le nombre de réseaux avec des poids différents, la probabilité de croisement et de mutation et le nombre de générations.

Jusqu'ici nous avons assumé que la structure du réseau est fixée et l'apprentissage génétique est employé seulement pour optimiser les poids dans le réseau donné.

II.3.3 Conclusion :

Les réseaux de neurones peuvent modifier les poids synaptique en faisant appel aux algorithmes génétiques afin d'éviter les minimum locaux.

Chapitre III

Résultats et interprétations

III. Résultats et interprétations

III.1 introduction :

Ce chapitre est consacré à l'implémentation de deux techniques dites intelligentes pour la reconnaissance du diabète.

Nous avons utilisé la base de données (*Pima Indian Diabetes*) pour classer les personnes en bonne santé et les personnes diabétiques dans deux classes différentes. Cette base est une collection de 768 exemples avec huit descripteurs et deux classes (normal et diabétique).

La première technique concerne des réseaux à bases radiales qui sont les réseaux probabilistes. Ce réseau a été développé par Specht en 1990. C'est une formulation de l'estimation de la densité de probabilité. C'est un modèle basé sur l'apprentissage compétitif avec ' un gagnant prend toute l'attitude '. Ce réseau est appelé probabiliste, il est vu comme l'implémentation de la méthode des fenêtres de Parzen, qui consiste à centrer une gaussienne sur chaque exemple d'apprentissage. Les estimateurs de Parzen ont été développés pour construire des fonctions de densité de probabilité qui sont exigées par la théorie de Baies.

La deuxième est une hybridation des réseaux de neurones avec les algorithmes génétiques. Nous avons utilisé les réseaux neuronaux artificiels comme l'une des méthodes puissantes dans le domaine d'intelligence artificiel pour classer les patients diabétiques en deux classes. Pour obtenir de meilleurs résultats, nous utilisons un algorithme génétique pour la sélection des paramètres pertinents du diabète. Après cela, ces fonctions sélectionnées ont été appliquée au réseau de neurone probabiliste.

A la fin, nous avons fait une étude comparative entre les deux techniques.

Dans ce chapitre nous réalisons deux classifieurs différents pour la reconnaissance du diabète:

- Un classifieur neuronal probabiliste (CNP)
- Un classifieur neuro-génétique (CNG)

III.2 Etat de l'art :

Plusieurs de travaux dans l'état de l'art ont traité la reconnaissance et la classification des types du diabète, ainsi que l'identification des différents facteurs qui sont à l'origine du diabète comme le facteur d'obésité, les facteurs génétiques, etc.
Quelques recherches dans la littérature avaient contribué pour améliorer la reconnaissance et l'identification de cette maladie, Parmi ces contributions nous citons :

- Vosoulipour et autres ont utilisé les algorithmes génétiques pour la sélection des paramètres pertinents du diabète, en choisissant l'ANFIS d'une part et un réseau de neurone multicouche d'autre part. Ils ont obtenu un taux de classification de 81.30 % pour l'ANFIS et 77.60% pour le réseau de neurones avec une base de test de 100 exemples. **[Vosoulipour 2008].**
- Dans le travail de **[Sharifi 2008]** ils ont testé plusieurs classifieurs avec huit et quatre descripteurs. Parmi ces classifieurs, nous citons le réseau MLP (multicouche) avec un taux de classification de 66.67% pour 8 descripteurs et un réseau à bases radiales (RBR) avec un taux de classification de 67.71% pour 8 descripteurs.
- Dans le travail de **[Sekkal 2009]** ils ont appliqué les algorithmes génétiques combinés avec les réseaux de neurones pour l'amélioration de l'architecture sur la base de données d'arythmies cardiaques de base MIT BIH, ils ont obtenu un taux de classification de 98.86%, une sensibilité de 99.09% et une spécificité de 98.66%. ils ont utilisé les algorithmes génétiques fondés sur un classifieur neuronal pour les poids de connexions sur la même base de données. Cette approche a donné de très bons résultats avec un taux de classification correcte de 98.72% et une sensibilité de 97.33% par rapport à classifieur classique qui a un taux de classification de 95.71% et 87.98% de sensibilité.
- Dans le travail de Zhang et autres, ils ont employé un algorithme génétique avec réseau de neurone pour la classification du cancer du sein avec un taux de reconnaissance de 90.5%. **[Zhang 2004].**

III.3 Problématique :

L'analyse classique (visuelle) des données médicales est devenue plus difficile du à l'augmentation et la diversité des facteurs de risque ainsi que les causes à l'origine des maladies. Il n'y a aucun doute que l'évaluation de données prises du patient et la décision de l'expert sont les facteurs les plus importants dans le diagnostic, donc la tache n'est pas aussi facile vu le nombre élevé des facteurs considérés ce qui justifie les travaux de recherche engagés actuellement dans plusieurs domaines médicales comme le diabète. Comme toute maladie la reconnaissance du diabète est basé sur beaucoup de facteurs , ce qui complique le diagnostic du médecin. Des systèmes de classification ont été développé pour analyser et identifier les données et extraire des informations pertinentes dans un temps plus court et d'une manière robuste, ce qui permet d'aider l'expert et éviter les erreurs causées par la fatigue ou dans le cas d'un expert non spécialiste.

III.4 Description de la base de données :

La base de données Pima diabétiques Indiens, donnés par Vincent Sigillito, est une collection de rapports médicaux de diagnostic de 768 exemples tirés d'une population vivant près de Phoenix, Arizona, USA. Cette base de donnée est composée par 268 femmes diabétiques et 500 femmes non diabétiques, l'âge de ces femmes varie ente 21 et 81 ans. Cette base de données est téléchargée à partir du site officiel de l'UCI (machine Learning repository).

Le but est de reconnaître la présence du diabète chez ces patients. Ces malades sont caractérisées par 8 attributs explicatifs :

Nombre	Attributs	Description
1	Prégnant	Nombre de fois où enceintes (Ngross)
2	Glucose	concentration plasmatique de glucose (Gly)
3	diastolique	La pression artérielle diastolique (mm Hg) (PAD)
4	Triceps	Indice d'obésité (épaisseur du pli cutané du triceps (mm)) (Epai)
5	Insuline	Taux d'insuline de 2 heures (mu U / ml) (INS)
6	BMI	indice de masse corporelle [poids en kg / (taille en m)^2] (IMC)
7	Diabetes	Diabète fonction pedigree (PED) (Indice d'antécédents familiaux pour le diabète)
8	Age	Âge (années)

Tableau III.3 : Nombre d'attributs de la base de données Pima Indian

Les individus diabétiques appartiennent à la classe "1" (test positif)

et les non diabétiques appartiennent à la classe "0" (test est négatif). Cette population possède l'un des taux les plus élevés d'incidence du diabète dans le monde. Une partie de ces données, contient des descripteurs nuls: taux de glucose dans le plasma, pression artérielle, épaisseur de la peau (triceps), indice de masse corporelle, insuline.

- 5 individus ont un taux de glucose dans le plasma = 0
- 11 individus ont un indice de masse corporelle = 0
- 35 individus ont une pression artérielle = 0
- 227 individus ont une épaisseur de la peau (triceps) = 0
- 374 individus ont une valeur d'insuline = 0

La nouvelle base corrigée contient 392 femmes avec 262 femmes non diabétiques et 130 femmes diabétiques. Dans le tableau **III.4** nous présentons quelques informations statistiques sur cette base de données.

Descripteur	MIN	MAX	La moyenne	La dérivation standard
1 (Ngross)	0	17	3.3010	3.20
2 (Gly)	56	198	122.6276	30.82
3 (PAD)	24	110	70.6633	12.48
4 (Epai)	7	63	29.1454	10.50
5 (INS)	14	846	156.0561	118.69
6 (IMC)	18.2	67.1	33.0862	7.01
7 (Ped)	0.085	2.42	0.5230	0.34
8 (Age)	21	81	30.8648	10.18

Tableau III.4 : Caractéristiques statistiques de la base de données Pima Indian

La figure **III.32** téléchargé à partir du site de l'université de Lyon (httppbil.univ-lyon) donne une idée sur la répartition des cas diabétiques et non diabétiques dans la base originale.

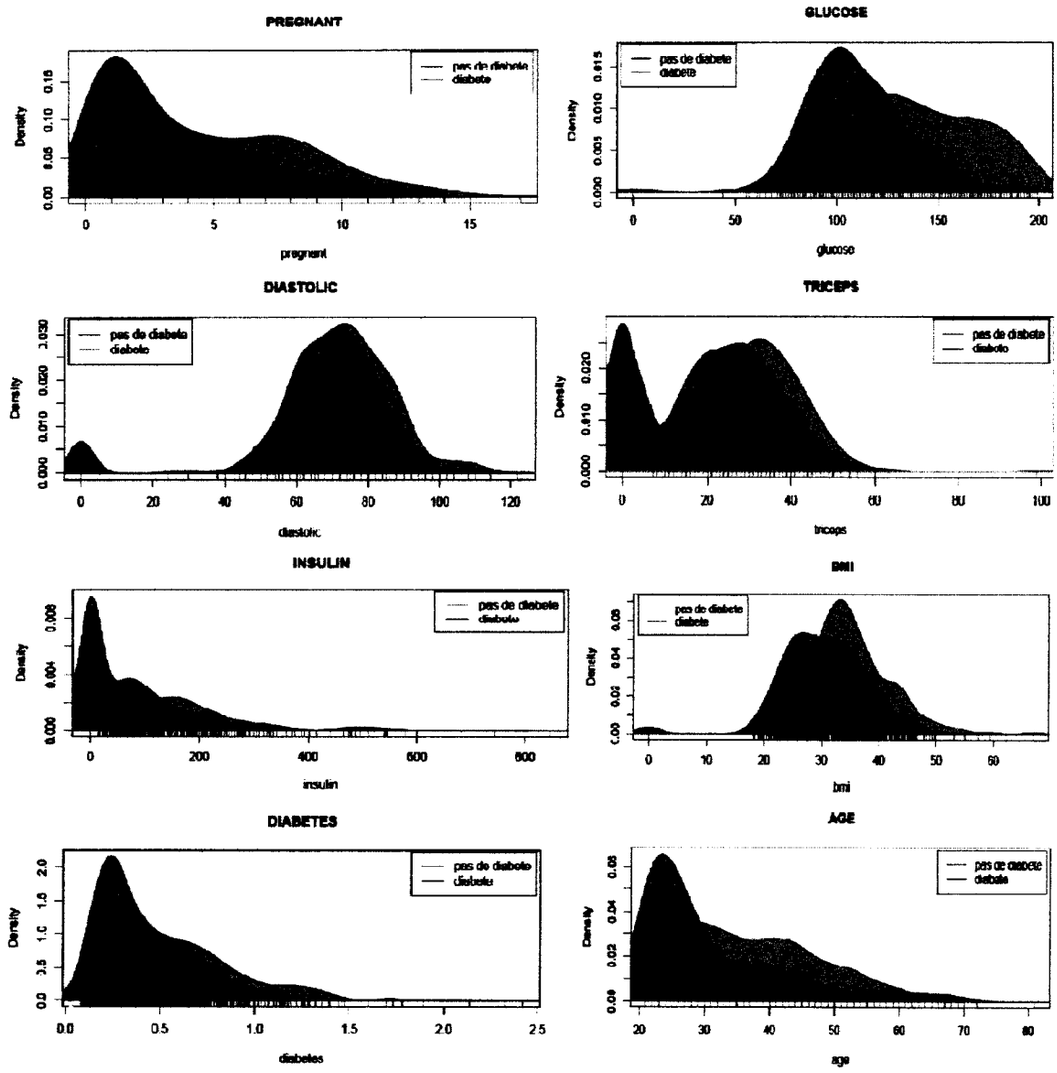


Figure III.32 : la représentions des descripteurs avant le filtrage de la base originale.

Les résultats obtenus après le filtrage des différentes valeurs nulles des descripteurs présentés dans la figure (**III.32**) sont illustrés dans la figure (**III.33**) **[LYON]**

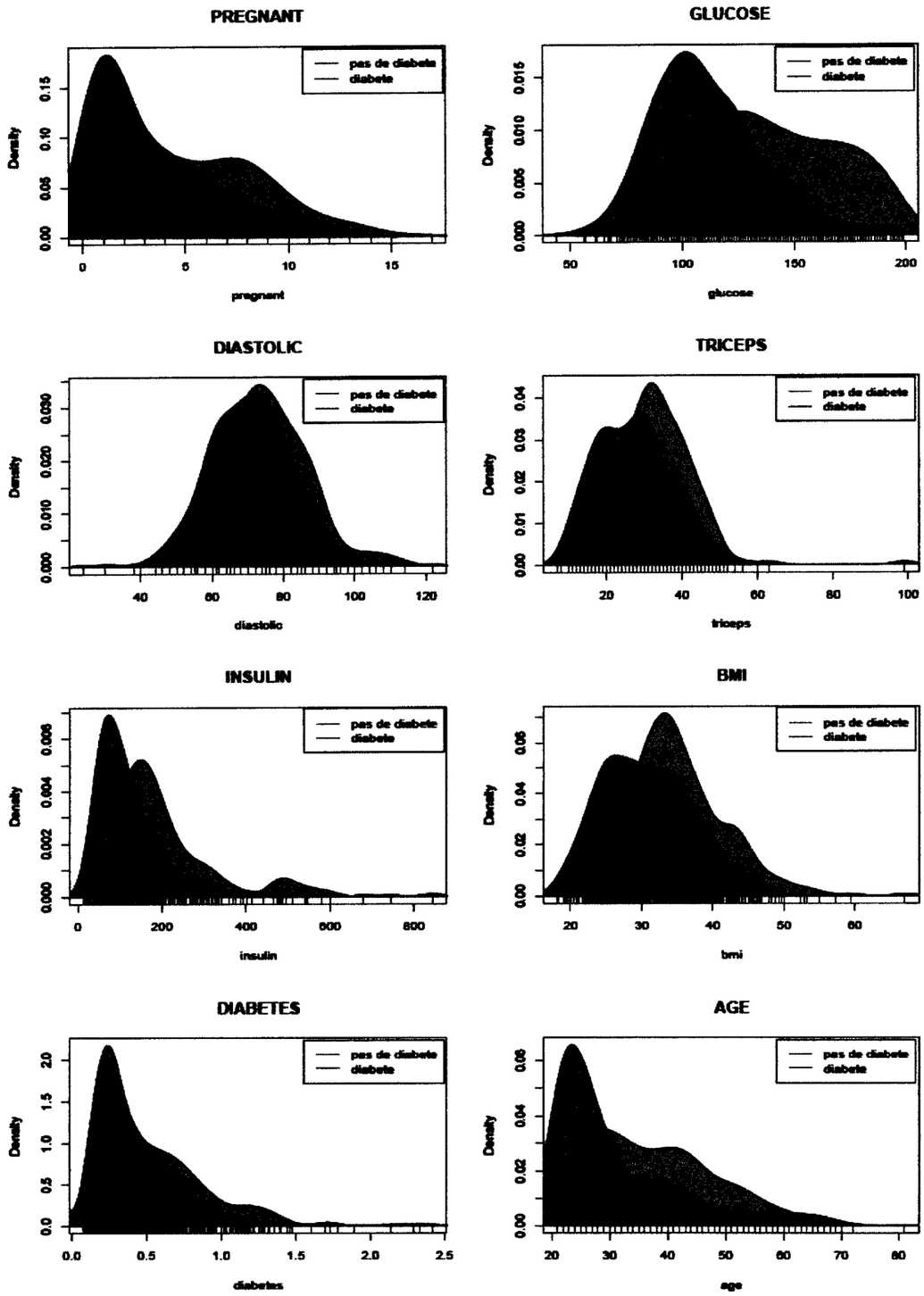


Figure III.33 : Les résultats obtenus après filtrage.

III.5 Corrélation entre les données : [Ammar2009]

Pour voir l'effet de chaque entrée sur la sortie (la classe) nous avons calculé la corrélation entre les différents attributs et la classe de sortie. Les résultats sont présentés dans le tableau III.5

Les coefficients de corrélation sont calculés à l'aide de la formule suivante :

$$r = \frac{\sum(xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum(xi - \bar{x})^2 (yi - \bar{y})^2}}$$

Colonne	Descripteurs	Corrélation (392 cas)
1	Prégnant-Nombre de grossesses. (Ngross)	0.2566
2	Concentration de glucose dans le plasma (Glu)	0.5157
3	pression artérielle diastolique (PAD)	0.1927
4	Triceps-Indice d'obésité (mm). (Epai)	0.2559
5	Taux d'Insuline (INS)	0.3014
6	Indice de masse corporelle (IMC)	0.2701
7	Fonction pedigree du diabète. (Ped)	0.2093
8	Age	0.3508

Tableau III.5 : Corrélation entre différentes entrées et la classe de sortie

La corrélation la plus élevée « 0.5175 » entre la classe et la concentration du glucose dans le plasma. Ce descripteur peut être d'un grand intérêt dans la prédiction du diabète. Cependant, il n'est pas toujours évident. Concernant la corrélation entre les différents attributs de caractérisation nous citons ceux ayant une corrélation mutuellement grande:

- Corrélation entre le nombre de grossesses et l'âge : 0.6796
- Corrélation entre l'épaisseur du pli cutané (triceps) et l'indice de masse corporelle : 0.66
- Corrélation entre concentration du glucose dans le plasma et le taux d'insuline : 0.58
- Corrélation entre La PAD et l'IMC : 0.30
- Corrélation entre La PAD et l'âge : 0.30

Ces résultats sont reportés dans le tableau **III.6** :

Descripteurs	Corrélation (392 cas)
l'âge et le nombre de grossesse	0.6796
l'indice de masse corporelle et l'épaisseur du triceps	0.66
Glycémie et l'insuline	0.58
pression diastolique et l'indice de masse corporelle	0.30
pression diastolique et l'âge	0.30

Tableau III.6 : Corrélation mutuelle entre certaines entrées

III.5.1 La corrélation entre le nombre de grossesses et l'âge :

Ces deux descripteurs ont une corrélation de 0.6796; ainsi dans la plus part des cas le nombre de grossesses est proportionnelle avec l'âge.

- Une boite à moustaches permet une représentation synthétique d'un ensemble des données : Les principes du diagramme boîte à moustaches **Figure III.34**

  - Une boîte dont les extrémités représentent les 25 et 75 pourcent.
  - la médiane de l'échantillon divise cette boîte en deux parties.
  - De part et d'autre des extrémités de la boîte, un trait (une moustache) horizontal représente le dernier des points de l'échantillon situé à une certaine distance (généralement 1.5 fois l'espace interquartile par défaut) de la boîte.
  - Les points situés au delà de cette limite sont identifiés par un symbole. Ils représentent les points « aberrants ».

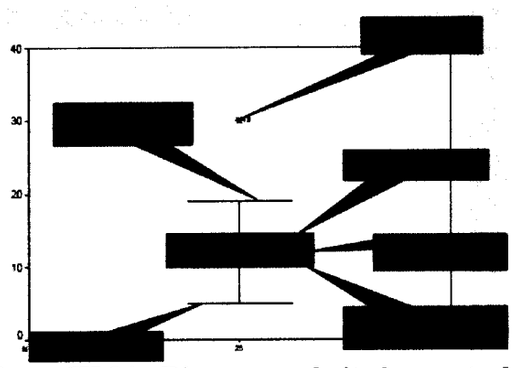


Figure III.34 : Diagramme boite à moustaches

- La corrélation entre le nombre de grossesses et l'âge est la plus élevé, Généralement la majorité des femmes ont un nombre de grossesses important en avançant dans l'âge.

  D'après la figure **III.35** nous remarquons deux scénarios:

  - **Les cas non diabétiques:** parmi les 262 cas non diabétiques totaux il y a 209 femmes ont un nombre de grossesses inférieur ou égal à 4 et 197 femmes ont un âge inférieur à 30 ans. Pour les 209 femmes nous trouvons 187 femmes avec un âge inférieur à 30 ans.
  - **Les cas diabétiques:** parmi les 130 cas diabétiques totaux il y a 111 femmes ont un nombre de grossesses supérieur ou égal à 1 et 114 femmes ont un âge supérieur à 25 ans. Pour les 111 femmes nous trouvons 98 femmes avec un âge supérieur à 25 ans.

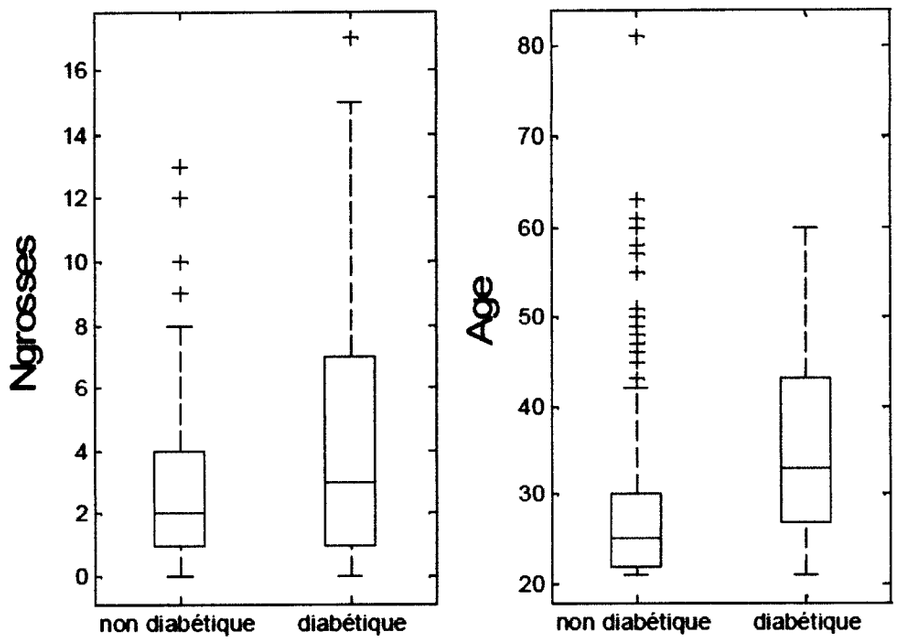


Figure III.35 : la représentation en boites à moustaches du nombre de grossesses et l'âge.

Nous constatons d'après ces résultats que la majorité des femmes non diabétiques sont jeunes
(Âge inférieur à 30 ans) avec peu de grossesses. Par contre la majorité des femmes diabétiques sont plus âgées avec plus de grossesses.

III.5.2 La corrélation entre la glycémie et l'insulinémie 2 heures (TOTG) :

Le glucose est une forme de glycémie que l'organisme transforme à partir du sucre et de l'amidon de la nourriture ingérée. Le corps l'utilise comme source d'énergie. On a tous besoin d'insuline pour décomposer la nourriture, ce qui explique la forte corrélation (0.58) entre ces deux descripteurs; ainsi plus le taux de glycémie est important et plus il ya une intolérance au glucose d'où plus de sécrétion d'insuline.

Dans la figure **III.36** nous remarquons deux scénarios:

- Les cas non diabétiques: parmi les 262 cas non diabétiques totaux il y a 211 femmes ont une glycémie inférieur à 130 mg/dl et 206 femmes ont un taux d'insuline inférieur à 180 mu U/ml. Pour les 211 femmes nous trouvons 180 femmes avec un âge inférieur à 180 mu U/ml.
- Les cas diabétiques: parmi les 130 cas diabétiques totaux il y a 84 femmes ont une glycémie supérieur ou égal à 130 mg/dl et 96 femmes ont un taux d'insuline supérieur à 130 mu U/ml. Pour les 84 femmes nous trouvons 70 femmes avec un taux d'insuline supérieur à 130 mu U/ml.

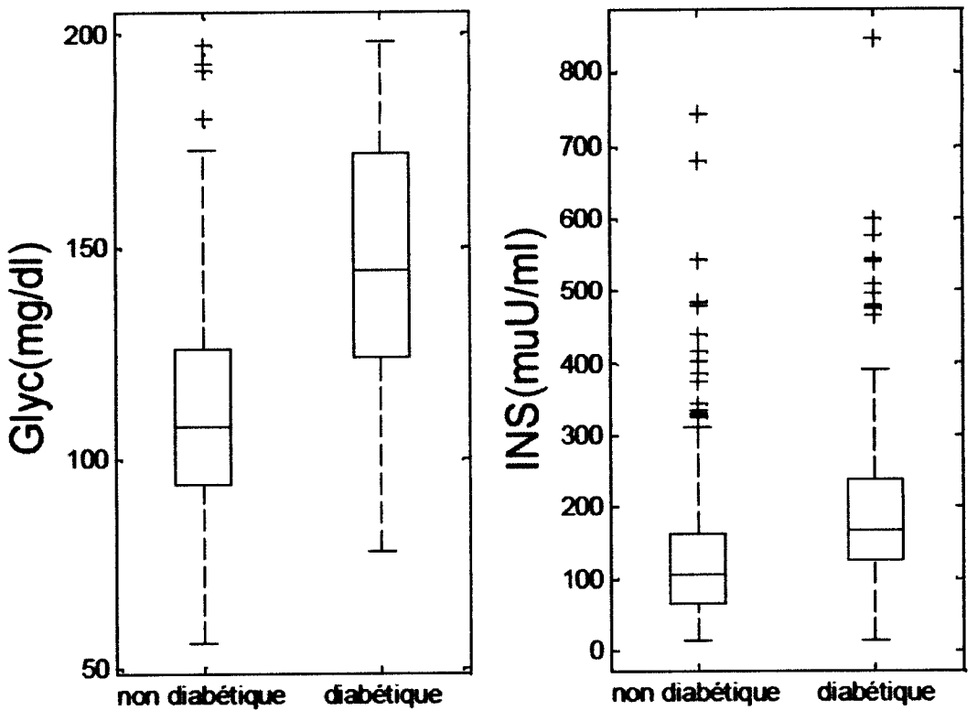


Figure III.36 : la représentation en boites à moustaches des paramètres: la glycémie et le taux d'insuline.

Nous remarquons que chaque fois la glycémie est petite le taux d'insuline est aussi petit pour les femmes non diabétiques. Par contre pour les femmes diabétiques une glycémie élevée est associé généralement à un taux d'insuline élevé.

III.5.3 La corrélation entre l'épaisseur de la peau au niveau du triceps et l'IMC :

Une mesure appelée indice de masse corporelle (IMC) ne prend pas directement les dimensions de la masse grasse, mais c'est un outil utile pour évaluer les risques pour la santé associés à une surcharge pondérale ou à l'obésité. Et l'épaisseur du triceps ou bien La graisse sous-cutanée représente approximativement 80 % du total de la graisse corporelle. L'épaisseur des plis adipeux sous-cutanés est donc une bonne estimation de la réserve calorique. Elle est souvent utilisée dans l'identification de l'obésité.

Ces deux descripteurs ont une corrélation de 0.66, elle est aussi un peu élevée; ainsi plus le poids est volumineux plus l'épaisseur du triceps qui est l'indice d'obésité est importante.

Dans la figure **III.37** nous remarquons deux scénarios:

- **Les cas non diabétiques:** parmi les 262 cas non diabétiques totaux 195 femmes ont un IMC inférieur à 36 kg/ m^2 et 194 femmes ont une épaisseur inférieure à 34 mm. Pour les 195 femmes nous trouvons 174 femmes avec une épaisseur inférieure à 34 mm.
- **Les cas diabétiques:** parmi les 130 cas diabétiques totaux 110 femmes ont un IMC supérieur à 30 kg/ m^2 et 101 femmes ont une épaisseur supérieure à 25 mm. Pour les 110 femmes nous trouvons 94 femmes avec une épaisseur supérieure à 25 mm.

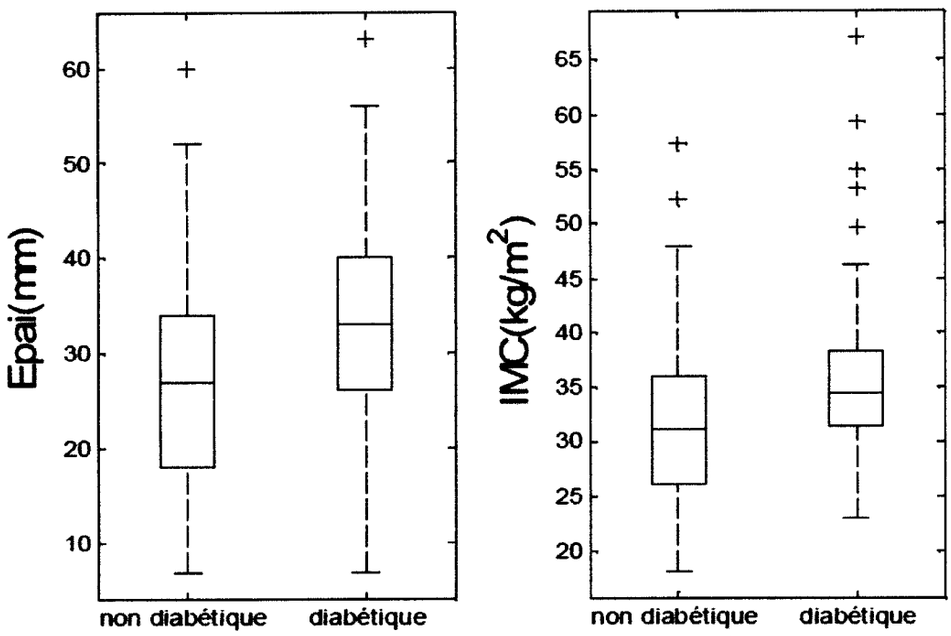


Figure III.37 : la représentation en boites à moustaches de l'épaisseur de la peau au niveau du triceps et l'IMC.

Nous constatons d'après ces résultats que la majorité des femmes diabétiques sont obèses avec une épaisseur de la plie cutanée grande. Cependant 50 % des femmes non diabétiques ont un IMC entre 25 et 36 kg/ m^2.

III.5.4 L'hérédité :

Elle est supposée comme un facteur de risque pour tous les types de diabète.
Dans la figure **III.38** nous remarquons deux scénarios:

- Les cas non diabétiques: parmi les 262 cas non diabétiques totaux il y a 192 femmes ont un Ped inférieur à 0.6.
- Les cas diabétiques: parmi les 130 cas diabétiques totaux il y a 103 femmes ont un Ped supérieur à 0.3.

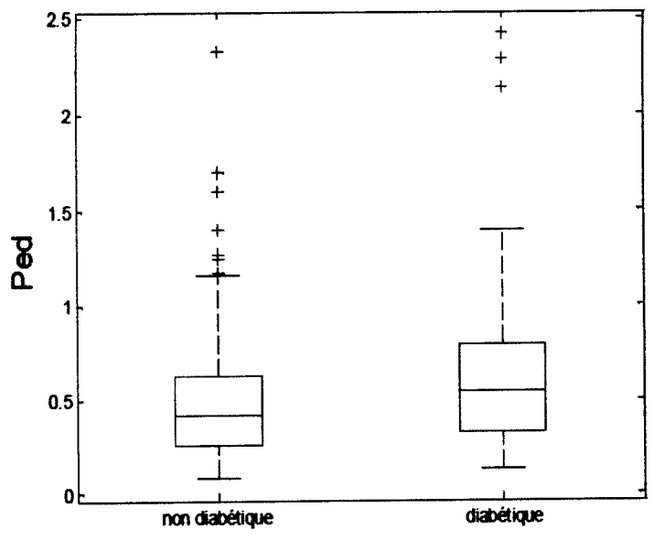


Figure III.38 : la représentation en boites à moustaches du pedigree du diabète

III.5.5 Hypertension artérielle :

Elle est associée généralement au diabète. Le seuil de risque est définit à partir de 130 mm Hg/85 mm Hg.

Malheureusement dans cette base les valeurs des cas diabétiques sont très proches de celles des cas non diabétiques ce qui rend ce paramètre non pertinent pour la classification.

Dans la figure **III.39** nous remarquons deux scénarios:

- Les cas non diabétiques: parmi les 262 cas non diabétiques totaux il y a 187 femmes ont une PAD inférieur à 75 mm Hg
- Les cas diabétiques: parmi les 130 cas diabétiques totaux il y a 100 femmes ont une PAD supérieur à 65 mm Hg.

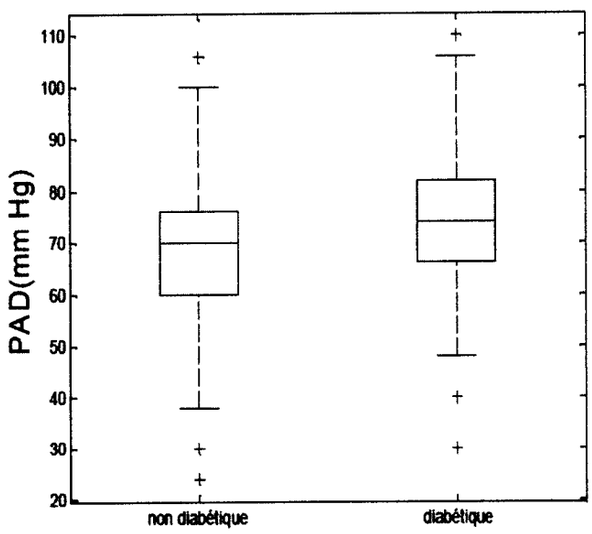


Figure III.39 : la représentation en boites à moustaches du pression artérielle diastolique.

III.6 Choix du langage de développement :

Après avoir défini clairement notre problématique, où nous avons expliqué tous les termes qui nous permettent de mieux cerner notre projet. Maintenant le choix du bon environnement de travail est très important, nous avons utilisés Matlab comme environnement de travail.

MATLAB **(MaTrix LaBoratory),** un produit de Mathworks, est un logiciel scientifique destiné à fournir des services intégrés de calcul numérique et la visualisation graphique en langage de programmation de haut niveau. Dr Cleve Moler, scientifique en chef de The Mathworks, Inc., a écrit à l'origine MATLAB, pour fournir un accès facile à des logiciels développés dans la matrice des projets LINPACK et EISPACK. La toute première version a été écrite dans les années 1970 pour utilisation dans des cours de théorie des matrices, algèbre linéaire et analyse numérique. MATLAB est donc construite sur un fondement de logiciels sophistiqués matrice, dans laquelle l'élément de base de données est une matrice qui ne nécessite pas de pré-dimensionnement.

MATLAB a une grande variété de fonctions utiles pour le praticien algorithme génétique et ceux qui souhaitent expérimenter avec l'algorithme génétique pour la première fois. Compte tenu de la polyvalence du langage **MATLAB** de haut niveau, des problèmes peuvent être codées dans M-Files dans une fraction du temps qu'il faudrait pour créer des programmes C ou Fortran dans le même but. Ajoutez à cela **MATLAB** analyse avancée de données, outils de visualisation et boîtes à outils spéciaux afin de domaine d'application et l'utilisateur est présenté avec un environnement uniforme pour explorer le potentiel des algorithmes génétiques.

III.7 Les modèles de classification :

La classification rassemble une famille de méthodes qui permettent d'automatiser le processus de reconnaissance. Le but de la classification est de classer les observations sur la base d'une serie de caractéristiques prédéfinies.

III.7.1 Chaine de classification des objets :

III.7.1.1 Les principaux modules de classification d'objet :

Un dispositif de classification ou de reconnaissance automatique de formes est généralement conçu comme une chaine de modules de traitement. Ainsi, un système de reconnaissance de formes comporte habituellement les modules suivants :

- **Un module de collecte des données :** Ce module permet de connecter les données nécessaire pour la classification on mesure en sortant nombre de caractéristique (la 1^{ére} représentation de l'objet).
- **Un module de prétraitement :** Ce module permet de faire le nettoyage des données et parfois réduire la représentation de la dimension de vecteur d'entrée (utiliser des outils comme l'algorithme génétique)
- **Un module de classification :** Ce dernier module est composé d'un algorithme de classification qui permet l'affectation d'un objet à ca classe respective.

Cet algorithme peut fournir soit une réponse binaire à valeurs discrètes (appartenance ou non à une classe) soit une réponse à valeurs continues.

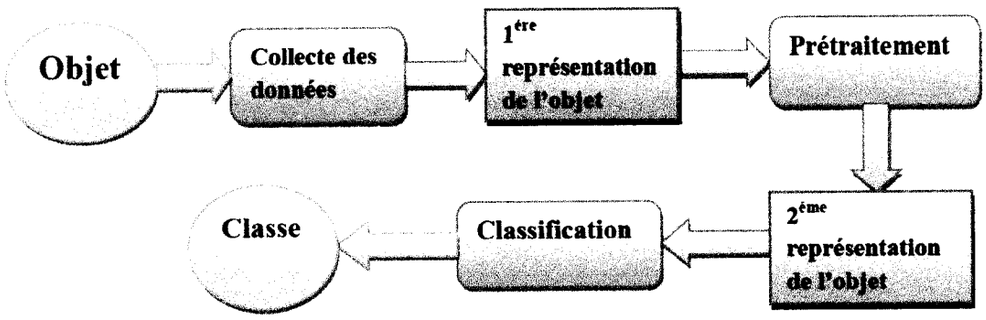


Figure III.40 : Chaine de module de classification

Remarque :

La tache de l'algorithme de classification est d'autant plus aisée que la représentation de l'objet est pertinente.

III.7.1.2 L'extraction des descripteurs :

Habituellement une classification de forme ne se fait pas directement sur des formes brutes, mais plutôt à partir de descripteurs ou paramètres caractérisant les formes. Il existe deux approches pour caractériser un objet quelconque.

- **L'approche purement mathématique :** qui applique des techniques et outils mathématique pour retenir un certain nombre de coefficient pertinente comme (analyse en composants principales, prédiction linéaire, transformer de Fourier, transformer par ondelettes,... etc.).
- **L'approche intuitive :** qui laisse au spécialiste le soin de définir les descripteurs qui lui semblent importants. Cette approche donne souvent de meilleurs résultats, car les paramètres choisis résultent. D'une grande expérience et peuvent être plus discriminants.

III.7.1.3 Procédure de résolution par apprentissage :

Le classifieur qui réalise le classement des objets doit passer par deux phase, une phase d'apprentissage et une phase de test.

- **Phase d'apprentissage :**

Le but de l'apprentissage est de découvrir les règles (généralement non déterministe) qui gouvernent et régissent des formes. L'apprentissage est un processus calculatoire qui doit être capable d'amener à une certaine prédiction et à une certaine généralisation. Il existe principalement trois types d'apprentissage : supervisé, non supervisé et renforcement. Dans le premier cas, le travail d'apprentissage consiste à analyser des ressemblances entre les formes d'une même classe et les dissemblances entre les formes de classes différentes pour déduire les meilleures séparateurs entre les classes ; dans le deuxième cas il s'agit de

regrouper les différentes forme en classe en fonction d'un critère de similarité choisit a priori « distance euclidienne » ; dans le troisième cas si les actions conduise le classifieur dans un état satisfaisant alors la tendance à produire ces actions doit être renforcer. **[CHIKH 2010]**

❖ **Phase de test :**

Cette phase doit permettre l'affectation d'un nouveau 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'utilisations. 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.

- **Taux de classification et taux d'erreurs :**

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 décrites dans le même espace de représentation que celles utilisées pour l'apprentissage. Elles sont aussi étiquetées par leur classe réelle d'appartenance 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 (la base de test ne doit avoir aucun objet en commun avec la base d'apprentissage et les éventuelles bases de validation). 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. Le découpage le plus courant est de **2/3** pour l'apprentissage et le **1/3** restant pour la base de test. En considérant que la base de test contient N objet et que sur ceux-ci N corrects sont biens classés parle système. **[Sekkal2009]**

⊕ **Le taux de classification (TC %):** définit l'ensemble des classes corectement classés par le classifieur,il représente le taux de classification totale.

$$TC(\%) = \frac{\textbf{N corrects}}{\textbf{N total}} * \mathbf{100}$$

Soit : N corrects= VN+VP ; N total= VP+VN+FP+FN

⊕ **Taux d'erreur :** représente l'ensemble des classes mal classés par le classifieur.

$$\mathbf{TE(\%)} = \frac{\textbf{N erreurs}}{\textbf{N total}} * \mathbf{100}$$

Soit : N erreurs= FP+FN ; N total= VP+VN+FP+FN

- **Sensibilité et spécificité :**

L'évaluation des performances d'un classifieur peut etre réalisée par l'appréciation de deux lois statistiques, qui sont la sensibilité et la spécificité.Pour rappel, ces deux quantités sont définies par :

- **La sensibilité (Se(i)):** ce paramètre représente le taux de classification des cas diabétiques (détectés correctement par le classifieur) par rapport au nombre total des cas diabétiques réels.

$$Se(i) = \frac{VP(i)}{(VP((i) + FN(i))}$$

- **La spécificité (Sp(i))** : la spécificité indique le taux de classification des cas non diabétiques (détectés correctement par le classifieur) par rapport au nombre total des cas non diabétiques réels.

$$sp(i) = \frac{VN(i)}{(VN(i) + FP(i))}$$

- **La précision (Prec(%)) :**

Représente le taux de classification des cas diabétiques correctement reconnus par le classifieur par rapport au nombre total des cas diabétiques détectés par le classifieur.

$$Prec(\%) = \frac{VP}{(VP + FP)} * 100$$

Ces paramètres sont calculés à l'aide de la matrice de confusion donnée par le tableau **III.7**
Avec:

- Vrai positif (VP(i)): un exemple diabétique prédit diabétique.
- Vrai négatif (VN(i)): un exemple non diabétique prédit non diabétique.
- Faux positif (FP(i)): un exemple non diabétique prédit diabétique.
- Faux négatif (FN(i)) : un exemple diabétique prédit non diabétique.

	Présence d'evenement de classe i	Absecence d'evenement de classe i
Classification positive	VP(i) :Vrai positif	FP(i) : Faux Positif
Classification négative	FN(i) :Faux Négatif	VN(i) : Vrai Négatif

Tableau III.7 : La matrice de confusion

III.8 Reconnaissance des diabètes :

Résoudre un problème de classification, c'est de trouver une hypothèse qui peut lier l'ensemble des objets à classer (caractérisés par des variables descriptives choisies) et l'ensemble des classes.

Dans cette expérimentation nous réalisons deux modèles de reconnaissance du diabète en se basant sur deux classifieur différents :

- ✓ Un classifieur neuronal probabiliste (CNP)
- ✓ Un classifieur neuro-génétique (CNG)

III.8.1 Reconnaissance du diabète par le classifieur neuronal probabiliste (CNP) :

III.8.1.1 Implémentation et apprentissage :

La figure **III.41** représente le réseau probabiliste que nous avons implémenté dans l'environnement du logiciel utilisé. Ce réseau est composé essentiellement de deux couches:

- ❖ La première couche est une couche de neurones à fonctions à base radiale (FBR) , ce qui signifie 2 choses :
  - ✓ Les neurones calculent la distance entre un vecteur d'entrée et leur vecteur poids
  - ✓ Leur fonction de transfert est une gaussienne
- ❖ La deuxième couche est une couche de compétition : elle prend les sorties de la couche à FRB, remplace la sortie maximale de cette couche par un 1 et les autres par un 0.

Cette méthode fonctionne en apprentissage supervisé, c'est à dire que la phase de l'apprentissage nécessite comme information un ensemble des vecteurs d'entrée associés à une sortie désirée du réseau.

Il est important de signaler que chaque classe est constituée par une somme de gaussiennes dont l'amplitude et l'étalement restent à déterminer. Dans notre travail, l'étalement est déterminé d'une manière empirique (spread).

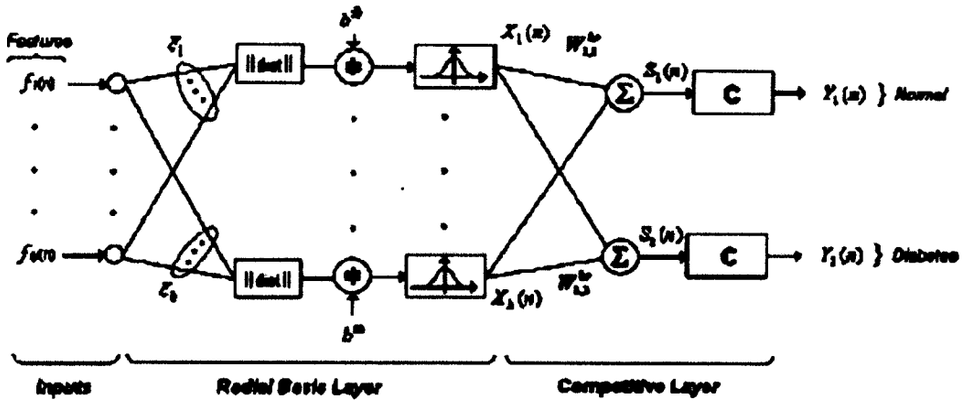


Figure III.41 : L'architecture du CNP implémenté

La phase d'apprentissage du réseau de neurones probabiliste est beaucoup plus facile que d'autre réseau de neurone classique. Le but de cette phase est de déterminer les poids entre la couche d'entrée et la couche cachée et aussi les poids entre la couche cachée et le couche de sortie.

Les poids entre la couche d'entrée et la couche cachée sont déterminés de la manière suivante:

$$\mathbf{wij = pij}$$

Où **wij** est le poids entre le neurone i de la couche d'entrée et le neurone j dans la couche cachée, **Pij** prendre la valeur de la **iéme** variable du vecteur d'entrée j.

Les poids entre la couche cachée et la couche de sortie sont déterminés de la manière suivante:

$$\mathbf{wik = Tjk}$$

Où **wjk** est le poids entre le neurone j dans La couche cachée et le neurone k de la couche de sortie, **Tjk** sont mis à 1 si la sortie du neurone j de la couche cachée appartient à la classe k, autrement il est mis au zéro.

Après l'étape de l'apprentissage, notre réseau peut être utilisé pour la prédiction ou bien la reconnaissance. La sortie de la couche cachée est donc calculée par fonction à base radiale suivante:

$$\boldsymbol{outj = \exp\left(-\left(\sum_{i=1}^{n}(xi - wij)^2 * \frac{0.8326}{SPREAD}\right)^2\right)}$$

Où **outj** est la sortie de neurone j dans la couche cachée, $\boldsymbol{xi}$ est la valeur de la variable de la base de test, SPREAD est le paramètre qui permet d'ajuster la largeur du gaussienne.

L'entrée de la couche de sortie est calculée par la formule suivante:

$$\boldsymbol{ink = \sum_{j=1}^{n} outj * wjk}$$

Où **in*k*** est l'entrée du neurone k de la couche de sortie. Si la sorte in*k* de neurone k de la couche cachée est grande par rapport au d'autre neurone de cette couche, la sortie du réseau est égale à la classe k.

III.8.1.2 Résultats expérimentaux :

Afin d'évaluer les performances du réseau implémenté, nous avons tout d'abord fixé le nombre des descripteurs à 8. Ensuite les performances sont étudiées en fonction des bases des données (nous changeons les bases d'exemples « 100, 130 et 170 exemples »). Cette évaluation est réalisée avec plusieurs valeurs du paramètre « spread ».

- **1ère expérimentation :**

La base de test est composée de 100 exemples dont 31 cas diabétique et 69 cas normaux. le reste des données (292 exemples) est utilisé dans l'apprentissage. Les résultats obtenus sont présentés dans le tableau III.8

Spread	VN Cas	FP Cas	FN Cas	VP Cas	TC (%)	PREC (%)	Se (%)	Sp (%)
0.10	60	9	10	21	81.00	70.00	67.74	86.96
0.12	59	10	9	22	81.00	68.75	70.79	85.51
0.15	57	12	12	19	76.00	61.29	61.29	82.61

Tableau III.8 : les performances du CNP avec une base de test contenant 100 cas.

- **2ème expérimentation :**

La base de test est composée de 130 exemples dont 39 cas diabétique et 91 cas normaux. le reste des données (262 exemples) est utilisé dans l'apprentissage. Les résultats obtenus sont présentés dans le tableau III.9

Spread	VN Cas	FP Cas	FN Cas	VP Cas	TC (%)	PREC (%)	Se (%)	Sp (%)
0.25	83	8	15	24	79.23	77.27	43.59	94.51
0.12	82	9	15	24	81.54	72.73	61.54	90.11
0.29	90	1	28	11	77.69	91.67	28.67	98.90

Tableau III.9 : performances du CNP avec une base de test de 130 cas

- **3ème expérimentation :**

La base de test est composée de 170 exemples dont 48 cas diabétique et 122 cas normaux. Le reste des données (222 exemples) est utilisé dans l'apprentissage. Les résultats obtenus sont présentés dans le tableau III.10

Spread	VN Cas	FP Cas	FN Cas	VP Cas	TC (%)	PREC (%)	Se (%)	Sp (%)
0.25	117	5	27	21	81.18	80.77	43.75	95.90
0.12	120	5	27	21	81.76	67.35	68.75	86.89

Tableau III.10 : performances du CNP avec une base de test de 170 cas

III.8.2 Reconnaissance du diabète par un classifieur neuro-génétique :

Le but de cette deuxième expérimentation et d'implémenter un modèle **neuro-génétique** interprétable pour une reconnaissance explicite du diabète. Nous avons utilisé les huit attributs pour faire une hiérarchie des descripteurs, cette comparaison faite par la technique des AGs pour trouver les descripteurs les plus pertinents.
Pour comparer les deux techniques, nous avons gardé la même base de test utilisée dans la troisième expérimentation (base d'apprentissage de 222 cas « 20% pour la validation et 80% pour l'apprentissage » et la base de test de170 cas).

III.8.2.1 Choix des paramètres de l'AG:

Pour exécuter l'algorithme génétique, il faut définir certains paramètres tels que : **la taille de la population, les probabilités de mutation et de croisement** et **le nombre de générations**. Il est difficile de les fixer ou de trouver les meilleurs avant l'exécution de l'algorithme. Donc la fixation de ces paramètres est une question d'essai, pour notre étude pour chaque valeur ou technique de ces paramètres, nous réalisons plusieurs phase d'apprentissage, puis nous calculons l'erreur quadratique (différence entre la sortie réelle et la sortie désirés), le taux de classification, la précision, la spécificité, la sensibilité pour chaque technique ou valeur.

- **Choix de la taille de population :**

Le choix de la taille de population dépond à la taille d'individu (le nombre de gène dans un chromosome). Dans notre cas les gènes sont des poids synaptiques c à d la taille de population dépond à la structure de réseau.
Nous avons réalisé un programme qui fait plusieurs apprentissages de différentes tailles de population : 20, 40, 60, 80, 100 individus.
Mais le choix de la taille diffère d'un classifieur à un autre (dépend de la structure), et la fixation d'une taille pour ce classifieur nous donne un repère pour les autres.

Taille	Erreur	TC	Prec	Se	Sp
20	**12**	**81.76**	**70.73**	**60.42**	**90.16**
40	**11**	**82.35**	**71.43**	**62.50**	**90.16**
60	**13**	**81.18**	**70**	**58.33**	**90.16**
80	**15**	**81.76**	**70.73**	**60.42**	**90.16**
100	**12**	**81.18**	**70**	**58.33**	**90.16**

Tableau III.11 : performances d'un classifieur CNG avec différentes taille de population

La taille de la population doit être choisie de façon à réaliser une erreur plus petite et un bon taux de classification, on remarque bien qu'une population de 40 individus réalise un taux de classification plus élève égale à 82.35% et l'erreur la plus petite.

Donc le bon choix de la taille de population est 40 individus.

➢ **Choix de la fonction scaling :**

La fonction Scaling ou mise à l'échelle permet de modifier la fonction fitness ; ils existent plusieurs techniques de scaling. Nous utilisons 2 types :

- ❖ Fonction Rank
- ❖ Fonction proportionnelle.

Ces 2 techniques sont présentées dans chapitre 2.

Types	Erreur	TC	Prec	Se	Sp
Rank	11	82.94	73.17	62.50	90.98
Prop	15	78.24	64.10	52.08	88.52

Tableau III.12 : performances du CNG avec différentes techniques de scaling

Ces résultats montrent bien que le classifieur neuro-génétique doit être plus performant pour les fonctions scaling de type « Rank ».

Donc la meilleure technique de scaling pour notre étude est la technique « **Rank** »

➢ **Choix de sélection :**

La sélection est le choix des parents pour la génération suivante en basant sur les valeurs de fitness scaling. Un individu peut être choisit plusieurs fois, par contre un autre peut être supprimé.
Dans cette étude, nous choisissons entre les 4 techniques de sélection : « stochastique », « remainder », « roulette », « tournoi ».

Sélection	Erreur	Tc	Prec	Se	Sp
Stochastique	12	80.59	69.23	56.25	90.16
remainder	13	78 .82	65.79	52.08	89.34
Roulette	16	81.76	70.73	60.42	90.16
Tournoi	15	82.94	74.36	60.42	91.80

Tableau III.13 : performances du CNG avec différentes technique de sélection

Nous remarquons que les technique « Tournoi »et «Roulette » donnent des meilleurs résultats par rapport aux techniques « Stochastique » et « remainder ».
Par exemple le taux de classification de tournoi égale à 82.94%, donc on peut assurer des bons résultats par rapport à l'autre technique.

Nous avons choisi pour notre application la technique « Tournoi »

➢ **Choix de taux de croisement :**

Taux de croisement est le taux d'individus qui participe à l'opérateur de croisement, pour que le reste participe à la mutation. Même expérimentation que les autres paramètres, nous avons réalisé plusieurs apprentissages de différent taux : 0, 0.2, 0.4, 0.6, 0.8, 1. Puis nous calculons Erreur quadratique, TC, Prec, Se et Sp.

Taux	Erreur	Tc	Prec	Se	Sp
0	12	80.00	68.42	54.17	90.16
0.2	15	77.65	63.89	47.92	89.34
0.4	13	78.82	65.79	52.08	89.34
0.6	13	78.82	65.79	52.08	89.34
0.8	10	82.94	74.36	60.42	91.80
1	11	81.18	73.53	52.08	92.62

Tableau III.14 : performance du CNG avec taux de croissement différents.

Les meilleurs résultats sont obtenus pour un taux égal à 0.8. Même dans la littérature plusieurs chercheurs ont trouvé que les meilleurs taux sont entre [0.6, 0.95]
Pour un taux égal à 0 (pas de croisement, tous les enfants sont des enfants de mutation), le classifieur peut avoir un **Tc** = 80%. La même chose pour un taux égale à 1 (pas de mutation toutes les enfants sont des enfants de croisement).
Ces résultats montrent bien l'importance de mutation et croisement dans les algorithmes génétiques.

***Nous vous fixé pour notre cas un taux à* 0.8**

➢ **Choix de meilleures techniques de croisement :**

Classiquement le croisement est la recombinaison entre deux parents pour avoir deux enfants pour la génération suivants. Le but de croisement est d'enrichir la diversité de la population.
Dans notre cas, nous utilisons 5 techniques de croisement : « croisement étendu », « croisement heuristique », « croisement arithmétique », « croisement à 2 points », « croisement à 1 point »

Croisement	Erreur	Tc	Prec	Se	Sp
inter médiate	11	82.94	73.17	62.50	90.98
heuristique	12	79.41	67.57	52.08	90.16
Arithmétique	13	77.65	63.89	47.92	89.34
2 points	14	78.24	64.10	52.08	88.52
1 point	11	77.65	65.63	43.75	90.98

Tableau III.15 : performance du CNG avec différents techniques de croissement

***Pour notre cas la seule technique qui a donné des résultats satisfaisants est* « inter médiate ».**

➢ **Choix de meilleures techniques de mutation :**

La mutation consiste à changer ou permuter les gènes de chromosome parent, pour avoir un chromosome enfant.
Dans notre étude, nous utilisons trois techniques « gaussien », « uniforme », « adaptatif ».

Mutation	Erreur	Tc	Prec	Se	Sp
uniforme	23	79.41	66.67	54.17	89.34
Gaussien	21	83.53	73.81	64.58	90.98
Adaptatif	27	80.00	67.50	56.25	89.34

Tableau III.16 : performance du CNG avec différents techniques de mutation

La technique choisie pour notre étude est **Gaussien**

III.8.2.2 Résultats expérimentaux:

Les paramètres les plus performants pour les algorithmes génétiques changent d'un problème à un autre, il y a pas une règle qui permet de fixer ces paramètres avant l'exécution de l'algorithme.

Dans notre étude après les expérimentations précédentes choisisses sont présentés dans le tableau suivant :

Paramètres	Technique ou valeur fixé
taille de population	**40**
Scaling	**Rank**
Sélection	**Tournoi**
Taux de croisement	**0.8**
Croisement	**inter médiate**
Mutation	**Gaussien**

Tableau III.17: différents paramètres choisie

➢ **Les poids des descripteurs selon les paramètres les plus performants :**

Descripteurs	Ngross	Gly	PAD	Epai	INS	IMC	Ped	Age
Les poids(%)	40.38	91.06	74.40	75.28	88.63	60.74	87.24	76.42

Tableau III.18 : Les poids synaptiques des descripteurs

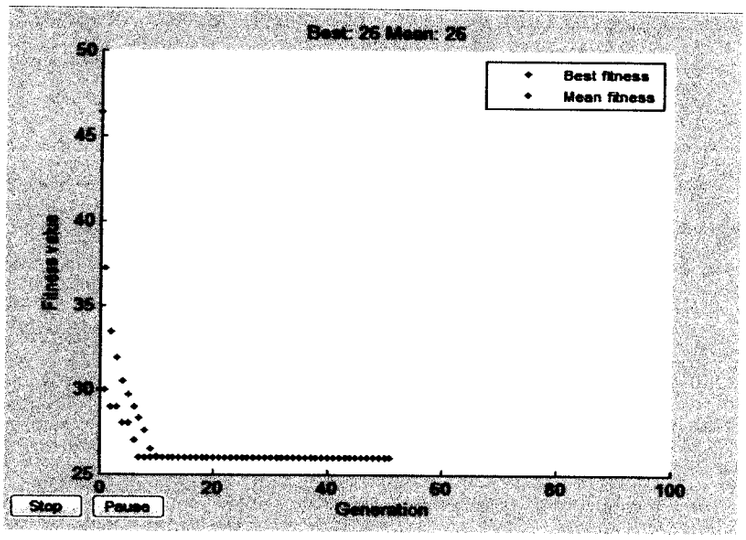


Figure III.42:variation des valeurs de fitness en fonction de génération

Résultats de classifieur neuro-génétique selon le nombre de génération :

Générations	Erreur optimal	TC (%)	Prec (%)	Se (%)	Sp(%)
100	**21**	**83.53**	**73.81**	**64.58**	**90.98**
80	29	80	67.50	56.25	89.34
40	22	81.18	70	58.33	90.16
20	29	81	69.50	58	90.16
50	23	81.18	70	58.33	90.16
60	23	82.35	71.43	62.50	90.16

Tableau III.19 : Taux de classification en fonction du nombre de génération

III.8.3 Interprétation des résultats et commentaire :

Nous remarquons d'après les résultats que :

- Le réseau de neurones probabiliste a donné de bons résultats. Cependant ces résultats restent non interprétables et cette caractéristique est très demandée même exigée dans le domaine médical.
- Pour le classifieur neuro-génétique, il a donné des meilleurs résultats et une amélioration remarquable par rapport au classifieur neuronal probabiliste.
- D'après le tableau **III.18** les descripteurs tel que taux de glucose dans le plasma et le taux d'insuline donnent des pourcentages plus élevés par rapport aux autres descripteurs qui nous considérons les plus pertinents et même les descripteurs tel que indice d'hérédité et âge sont aussi des facteurs de diabète. Ces résultats obtenus sont ainsi validés par l'expert humain (médecin).

III.9 Etude comparative :

Dans cette partie nous avons fait la comparaison entre les différents travaux existants dans la littérature et les résultats obtenus avec réseau probabiliste et neuro génétique, et cela sur la même base de donnée du Pima Indian.

Etude	Technique	Taux de classification
[Vosoulipour 2008]	AG-ANFIS	81.31%
[Vosoulipour 2008]	AG-PMC	77.60%
[Sharifi 2008]	PMC	66.67%
[Sharifi 2008]	FBR	64.06%
Notre travail	FBR	81.76%
Notre travail	AG-FBR	83.53%

Tableau III.20 : Etude comparative entre les différentes approches

III.10 Conclusion :

Par simple comparaison de nos résultats on conclue que la reconnaissance par un classifieur neuro-génétique et beaucoup plus satisfaisante par rapport au classifieur neuronal probabiliste.

Dans la méthode neuro- génétique, la difficulté du choix d'une bonne fonction de sélection et adaptation font des AGs une technique très particulière qui doit être bien maitrisée pour donner des bons résultats.

Dans le cas contraire, ces algorithmes n'arrivent pas à trouver de bonnes solutions aux problèmes posés.

Les AGs sont des algorithmes très lourds à exécuter, d'où l'intérêt d'exploiter le parallélisme dans ce type d'approche.

Conclusion générale et perspectives

Le diabète reste parmi l'une des maladies fréquemment rencontrées. De fait, plusieurs travaux ont été consacrés pour mieux comprendre son mécanisme, ses causes, sa reconnaissance et sa classification. En effet, ces deux dernières décennies, la reconnaissance et la classification du diabète continue d'être parmi les applications les plus souvent rencontrées dans le domaine médical. Dans ce travail de mémoire de master nous avons abordé en détail cette problématique.

La classification du diabète est basée sur deux approches intelligentes que nous avons implémentées. La première approche s'articule sur le réseau de neurones probabiliste alors que la deuxième utilise une approche neuro-génétique.

Selon les résultats obtenus, les réseaux de neurones probabilistes à bases radiales ont prouvé leurs capacités de résoudre les problèmes de classification, ceci confirme certains travaux existants déjà dans la littérature.

Concernant la deuxième approche, nous avons montré l'intérêt et l'interprétabilité des résultats et en a perspective une amélioration par rapport a la première approche et même l'expert humain arrive facilement à comprendre et interpréter les résultats obtenus.

REFERENCES BIBLIOGRAPHIQUES

A

[Ammar 2009] Mr. Ammar Mohammed, Reconnaissance Automatique Du Diabète Et Prédiction de la dose d'insuline, Mémoire de Magister en Electronique Biomédicale, Universté ABOU BAKR BELKAID-Tlemcen, promotion 2008-2009.

[AGGN] Site internet, http://www.a525g.com/intelligence-artificielle/algorithme-genetique.htm

[AVANT]
http://theses.univbatna.dz/index.php?option=com_docman&task=doc_download&gid=348&Itemid=4

[AGIA] Site internet, http://www.a525g.com/intelligence-artificielle/algorithme-genetique.php

B

[BIODIAB]http://www.espace-bio-millau.com/pages/diabete/le-diabete-c-est-quoi-symptomes-prevention-traitements-medicaux-etc-consulter-chacune-des-fiches-qui-leur-sont-consacrees.html

[BENDI 2008] BENDIMRED. I, FANDI. N, Classification des pathologies cardiaques par neuro-génétique, Mémoire d'ingenieur en Electronique Biomédicale, Universté ABOU BAKR BELKAID-Tlemcen,2008

C

[CHIKH 2010] Mr CHIKH Mohammed El Amine, Cour sur les réseaux de neurones. Master2 (SIC), Département d'informatique Faculté des sciences, 2010.

[CONCLUAG] http://www.ambafrancema.org/efmaroc/fontaine/college/IDD2006/iddg3.htm

[COMPDIA] Site internet, http://cprv.pagesperso-orange.fr/diabete.htm

D

[DIABHIS] Site internet, http://www.news-medical.net/health/History-of-Diabetes-(French).aspx,

[DIAB6] Site internet, http://www.bdaa.ca/biblio/apprenti/diabete/page6.htm,

[DIABGDM] Site internet, http://www.paperblog.fr/1456125/diabete-gestationnel/, 3 janv. 2009

[DIABSY] http://www.diabete.qc.ca/html/le_diabete/questcequedia.html, March 2001

[DIAB15] Site internet, http://www.bdaa.ca/biblio/apprenti/diabete/page15.htm

[DIAB24] Site internet, http://www.bdaa.ca/biblio/apprenti/diabete/page24.htm

[DIAB28] Site internet, http://www.bdaa.ca/biblio/apprenti/diabete/page28.htm

L

[LHABIT] Site internet, http://recherche.enac.fr/opti/papers/thesis/HABIT/main002.html

[LYON] Site internet, http://pbil.univ-lyon1.fr/R/enseignement.html

[LMSGY] http://www.parlonsdiabete.com/Mesure-du-glucose-en-continu.html,Decembre 2008

[LPATHO] Site internet, http://www.toutapprendre.com/minicours.asp?sante-bien-etre,pathologies,le-diabete-c-est-quoi&4542&1

R

[RN REC] Site internet, http://www.seas.upenn.edu/~timothee/papers/vision_system.pdf

[RN WIKI] Site internet, http://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels

[RN ing] Site internet, http://www-igm.univ-mlv.fr/~dr/XPOSE2002/Neurones/index.php

[RN course] Site internet, http://www.peoi.org/Courses/Coursesfr/neural/EL150FR.html

[RN COUR] Site internet,http://www.peoi.org/Courses/Coursesfr/neural/EL150FR.html

S

[SUCREDIA] Siteinternet,http://www.questmachine.org/article/Diab%C3%A8te_sucr%C3%A9

[Sekkal2009] Sekkal M, Apprentissage génétique d'un classifieur neuronal application en cardiologie, Mémoire de Magister en Electronique Biomédicale, Universté ABOU BAKR BELKAID-Tlemcen, 2009

[Sharifi 2008] Sharifi Arash, Vosolipour Asiyeh, *Hierarchical Takagi-Sugeno Type Fuzzy System for Diabete Mellitus Forecasting.*Computer department of Islamic Azad University Science and Research Branch, Tehran, Iran, 12-15 july 2008.

[SCHEMA] site Internet, http://e-philo.univ-paris1.fr/Smolensky1.htm

T

[TITI] H.TEBIB et Z.TITI, Modélisation par les réseaux de neurones optimisés par les algorithmes génétiques.PFE, université de M'sila, 2009

[TROAG] http://wcours.gel.ulaval.ca/2007/h/19968/default/5notes/Intro-AG.pdf

V

[Vosoulipour 2008] Vosoulipour A, M. Teshnehlab, *Classification on Diabetes Mellitus Data-set Based-on Artificial Neural Networks and ANFIS*, University of Technology/ Faculty of Electrical Engineering / Tehran/Iran, 2008

W

[WIKIAG] http://fr.wikipedia.org/wiki/Algorithme_g%C3%A9n%C3%A9tique

[WIKI2] http://fr.wikipedia.org/wiki/Diab%C3%A8te

Z

[Zhang 2004] Zhangn P, Verma B, Kumar K , A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography,2004.

Annexe A
Les réseaux à base radiales (Radial Basis Network)

Description :

Les réseaux à bases radiales (RBR) sont des modèles connexionnistes simples à mettre en œuvre et assez intelligibles, et sont très utilisés pour la régression et la discrimination. Leurs propriétés théoriques et pratiques ont été étudiées en détail depuis la fin des années 80 ; il s'agit certainement, avec le Perceptron multicouche, du modèle connexionniste le mieux connu. Une fonction de base radiale (RBF) est une fonction $\acute{A}$ symétrique autour d'un centre :

$$\mu_j : \phi_j(x) = \phi(\|x - \mu_j\|)$$

Par exemple, la fonction gaussienne est une fonction à base radiale (FBR) avec la norme euclidienne suivante: $\phi(r) = e^{-r^2/2\sigma^2}$

En général, les FBRs sont paramétrées par σ_j qui correspond à la « largeur » de la fonction

$$\phi_j(x) = \phi(\|x - \mu_j\|.\sigma_j)$$

Les différents types de réseaux de neurones à bases radiales de (Radial Basis Network)

La fonction de transfert est une exponentielle. L'opérateur sommation disparaît au profit de l'opération multiplication (élément par élément des matrices). Les réseaux à bases radiales nécessitent beaucoup plus de neurones qu'un réseau feedforward.

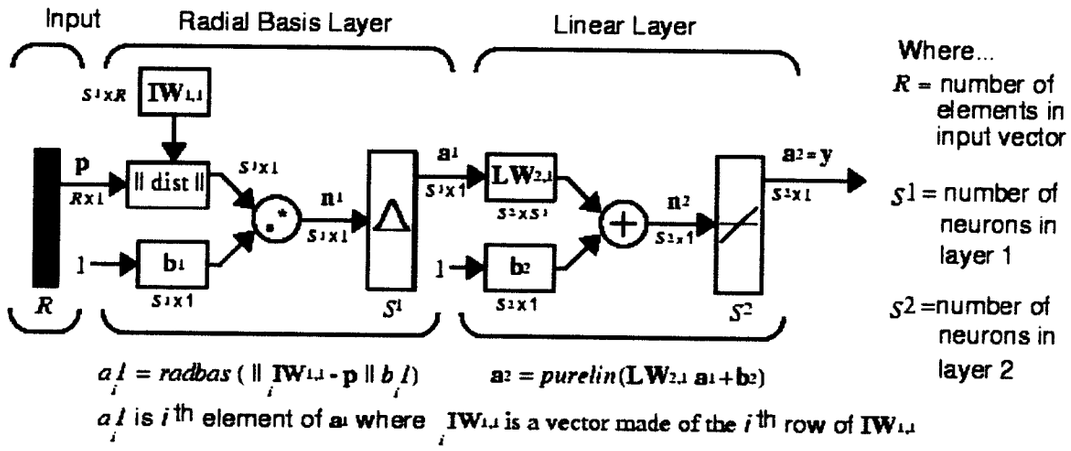


Figure A.43 : Architecture du réseau à fonction radiale

➢ **Les réseaux de neurones à régression généralisée (Generalized Regression Networks)**

Dans un réseau à régression généralisée (GRNN), il y a un réseau à base radiale auquel on ajoute une couche de sortie constituée d'une fonction de transfert linéaire. Nprod signifie une multiplication élément par élément, moralisé par la somme des éléments de a. Ces réseaux sont aussi utilisés en tant qu'approximation de fonction, mais sont plus lourds d'utilisation que les perceptrons multicouches.

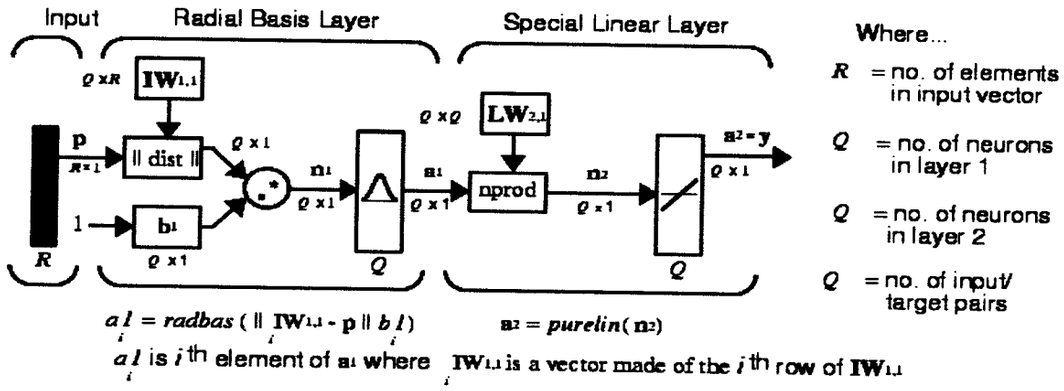


Figure A.44 : Architecture du réseau GRNN

➢ **Les réseaux de neurones probabilistes (Probabilistic Neural Networks)**

Ces réseaux sont généralement utilisés pour des problèmes de classification. La première couche qui est un réseau à base radiale, donne une information sur la ressemblance entre la donnée d'entrée et le jeu de données utilisé lors de l'apprentissage. La deuxième couche produit comme sortie un vecteur de probabilité. Finalement, une fonction de transfert compétitive produit 1 ou 0.

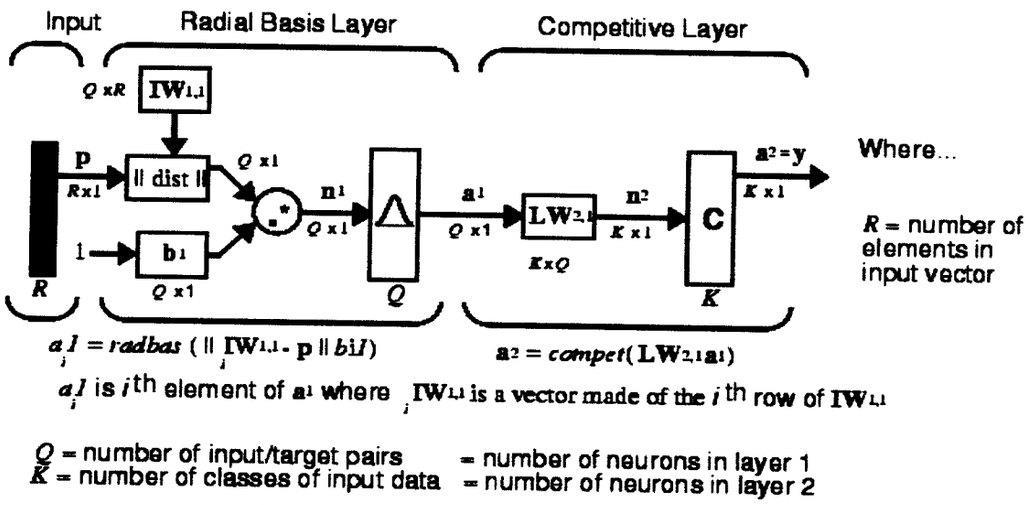


Figure A.45: Architecture du réseau PNN

Annexe B

EXPLICATIONS DES ÉLÉMENTS

Acidocétose :

Augmentation de corps cétoniques dans l'organisme conduisant à une acidose avec ou sans coma.

Anticorps :

Protéine formée par le système immunitaire pour se protéger de substances étrangères comme des bactéries ou des protéines étrangères.

Cellule alpha (α) :

Type de cellule contenue dans les îlots de Langerhans du pancréas et responsable de la production du glucagon, hormone hyperglycémiant.

Cellule béta (ß) :

Type de cellule contenue dans les îlots de Langerhans du pancréas et responsable de la production d'insuline.

Cétose, céto-acidose :

Accumulation de corps cétoniques provenant de la combustion des graisses lorsque l'organisme ne peut plus utiliser les sucres par manque d'insuline. La cétose conduit à une acidose avec ou sans coma.

Gène :

Unité de transmission du patrimoine héréditaire, responsable des caractères de l'individu. Chaque gène est spécifique d'un caractère précis, fait partie d'une portion de l'ADN portée par les chromosomes.

Glucagon :

Hormone hyperglycémiant, produite par les cellules alpha (α) du pancréas. Peut être utilisé par voie injectable dans le traitement d'une hypoglycémie sévère.

Glucides (ou hydrates de carbone) :

Une des trois classes principales d'aliments et principale source d'énergie de l'organisme, comprenant principalement les sucres et les amidons.

Glucose :

Sucre simple composé d'une seule molécule. Il est la principale source d'énergie de l'organisme. Pour que le glucose puisse être "brûlé" dans la plupart des tissus, une sécrétion appropriée d'insuline est indispensable.

Hormone :

Substance fabriquée par une glande et déversée dans le sang pour agir sur divers tissus. L'insuline est une hormone comme l'hormone de croissance fabriquée par l'hypophyse ou les hormones sexuelles fabriquées par les testicules ou les ovaires.

Index (ou indice) de masse corporelle :

Encore appelé BMI (body mass index) ou, en français, indice de Quételet, ce calcul repose sur la taille et le poids. La formule est : IMC = P/T2. P = poids en Kg ; T = taille en cm. La normale est toujours inférieure à 25 chez l'homme et habituellement inférieure à 23 chez la femme.

Insuline :

Hormone hypoglycémiante produite par les cellules béta (ß) des îlots de Langerhans du pancréas. L'absence d'insuline doit être compensée par l'injection régulière d'insuline, en fonction des glycémies.

Insulino-résistance :

Se caractérise par une réponse insuffisante des tissus à l'insuline. Fréquente chez l'obèse, elle peut occasionner une intolérance au glucose ou un véritable diabète.

Intolérance au glucose :

Etat caractérisé par une élévation modérée de la glycémie à jeun ou de la glycémie 2 heures après ingestion de glucose. Se voit surtout dans l'obésité "androïde" (à répartition mâle) : la graisse est sur le ventre.

Langerhans (Paul) :

Étudiant en médecine ayant fait sa thèse sur des amas de cellules (îlots) dans le pancréas endocrine. Il s'avéra plusieurs dizaines d'années plus tard que les cellules contenues dans ces "îlots" fabriquaient l'insuline.

Lipides (ou graisses) :

Une des trois classes d'aliments avec les sucres et les protéines. Forme de réserve énergétique principale dans les cellules adipeuses (graisseuses) sous forme de triglycérides. Leur combustion forme des acides gras, qui sont normalement source d'énergie. L'accumulation anormale des acides gras (en l'absence d'insuline) entraîne une acidose.

Pancréas :

Organe situé profondément derrière l'estomac. Comprend une partie endocrine qui fabrique de nombreuses hormones dont l'insuline et le glucagon, une partie exocrine qui secrète des enzymes favorisant la digestion.

Résumé :

Le diabète est une maladie chronique qui nécessite un soin médical continu avec une autogestion par le diabétique lui-même, pour éviter les complications à court terme (hypoglycémie, hyperglycémie...) et réduire le risque des complications à long terme (complications cardiaques, insuffisance rénale, rétinopathie, lésions nerveuses, endommagement des vaisseaux sanguins).

Dans ce mémoire, nous allons appliquer deux méthodes pour étudier la reconnaissance et la classification du diabète sur les données médicales de la base de données UCI (Machine Learning Data base).
Nous avons utilisé les réseaux neuronaux probabilistes comme l'une des méthodes puissantes dans le domaine de l'intelligence artificiel afin de classer les patients diabétiques en deux classes. Nous avons utilisé un algorithme génétique pour la hiérarchisation des descripteurs et de l'appliquer au réseau de neurones probabilistes.
Ces résultats obtenus indiquent que les méthodes proposées pour la reconnaissance du diabète sont très prometteuses.

Mots-clés : Classification, Diabète - Réseaux de Neurones - Algorithme Génétique, neuro-génétique.

Abstract:

Diabetes is a chronic disease that requires ongoing medical care with the diabetes self-management itself, to avoid short-term complications (hypoglycemia, hyperglycemia...) and reduce the risk of long-term complications (cardiac complications, renal failure, retinopathy, nerve damage, damage to blood vessels).
In this Work, we applied two methods to study the recognition and classification of diabetes on a medical data called UCI (Machine Learning Data Base).
We used probabilistic neural networks as one of the powerful methods in the field of artificial intelligence to classify diabetic patients into two classes. We used a genetic algorithm for the prioritization of descriptors and apply the probabilistic neural network.
These results indicate that the proposed methods for the recognition of diabetes are very promising.

Keywords: Classification, Diabetes - Neural Networks - Genetic Algorithm, neuro-genetics.

تلخيص:

داء السكري عبارة عن مرض مزمن يتطلب رعاية طبية مستمرة من طرف المريض نفسه لتجنب المضاعفات و التعقيدات على المدى القصير (نقص السكر في الدم، ارتفاع السكر في الدم...) والتقليل من خطر مضاعفات على المدى الطويل (أمراض القلب، القصور الكلوي، فقدان البصر، تضرر الأعصاب وتلف الأوعية الدموية).
في هذه المذكرة، طبقنا طريقتين لدراسة التصنيف و التعرف على مرض السكري على البيانات الطبية من قاعدة بيانات الجهاز التعليمي.استخدمنا الشبكات العصبية الاحتمالية باعتبارها واحدا من الأساليب القوية في مجال الذكاء الاصطناعي لتصنيف مرضى السكري إلى فئتين.استخدمنا خوارزمية الجينية لتحديد وتصنيف المعلومات الملائمة بمرض السكري و تم تطبيقها على الشبكة العصبية الاحتمالية.

النتائج المحصلة عليها تشير إلى أن الطرق المقترحة لتعرف على داء السكري جيدة.

الكلمات الرئيسية : تصنيف -- السكري -- الشبكات العصبية -- الخوارزميات الوراثية وعلم الجينات العصبية.