Les réseaux de neurones un outil de sélection de variables : Le cas des facteurs de risque de la maladie du cancer du sein

By/Par **Manel Zribi, Younes Boujelbene**

Faculté des Sciences Economiques et de Gestion, Université de Sfax, Tunisie
Unité de Recherche en Economie Appliqué

ABSTRACT

This paper uses the neural networks with an incremental learning algorithm as tool of selection of the most relevant risk factors in the disease of the breast cancer diagnosis. The results testify of the relevance of the neuronal approach with an incremental algorithm in research. An experimental study is performed using simulated data are further validated on real clinical data for breast cancer diagnosis; it was possible to us to determine the optimal combination of factors allowing to affect a good predictive performance of the type of tumor.

Keywords: Neural networks, incremental learning algorithm, classifications risk factors.

RÉSUMÉ

Ce papier utilise les réseaux de neurones avec un algorithme incrémental comme outil de sélection des facteurs de risques les plus pertinents dans la maladie du cancer du sein. Les résultats témoignent de la pertinence de l'approche neuronale avec un algorithme incrémentale dans ce domaine de recherche. A partir d'un échantillon de 248 patientes atteintes par cette maladie, il nous a été possible de déterminer la combinaison optimale des facteurs permettant d'atteindre une bonne performance prédictive du type de tumeur maligne et bénigne.

Mots clés : réseaux de neurones, algorithme incrémental, classification des facteurs de risques.

JEL Classification: C38

INTRODUCTION

Depuis une dizaine d'années, l'utilisation de réseaux de neurones artificiels (RNA) s'est développée dans de nombreuses disciplines (sciences économiques, écologie, environnement, biologie et médecine…). Ils sont notamment appliqués pour résoudre des problèmes de classification, de prédiction, de catégorisation, d'optimisation, de reconnaissance des formes et de mémoire associative (Drew et Monson, 2000).

Dans le cadre de traitement de données, les RNA constituent une méthode d'approximation de système complexes, grâce à leur capacité d'être un approximateur universel, ont prouvé leur capacité à extraire de données d'expérimentation des modèles performants, sans avoir à effectuer d'hypothèse sur la forme générale de ces derniers (Thomas Pet G.Bloch, 1999).

Avec les méthodes statistiques traditionnelles, il faut d'abord penser à un modèle, le tester, penser à un autre jusqu'à ce qu'on obtienne un modèle suffisamment précis. Un réseau est entrainé sur des données grâce à un mécanisme d'apprentissage qui agit sur les constituants du réseau pour réaliser au mieux la tâche désirée. La famille de RNA la plus utilisée ces dernières années comme outil d'aide à la décision est le perceptron multi-couche (PMC).

Ce type de réseau a été appliqué dans l'aide à la décision médicale pour le traitement de données pour l'anthropologie (Benoit le Blanc et al., 2001), pour la prévision des fractus (Baxt, 1995), pour le diagnostic des pathologies pulmonaires (Patil et al., 1993), du diabète (Armoni,1998), des cancers (Han et al., 2001), de la maladie d'Alzheimer (Hamilton et al.,1997) ,etc.

Ce papier n'est pas destiné à détailler le fonctionnement des réseaux de neurones vu que de nombreux articles et ouvrages font partie (Davalo et Naim, 1990 ; Bourret et al., 1991 ; Abdi, 1994 ; Cross et al.,1995), mais nous souhaitons montrer l'intérêt prévisionnel de cet outil par la sélection des variables (Baxt, 1995).

LES RÉSEAUX DE NEURONES ARTIFICIELS

Les RNA sont des modèles d'entrées-sortie basés sur un caractère des neurones biologiques, le but initial de cette modélisation est de reproduire les capacités du cerveau humain à interpoler ou à classifier.

Selon (Haykin, 1994), un RNA est un processus distribué de manière massivement parallèle, qui a une propension naturelle à mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour utilisation. Il ressemble au cerveau en deux points :

- la connaissance est acquise à travers un processus d'apprentissage

- les poids des connections entre les neurones sont utilisés pour mémoriser la connaissance.

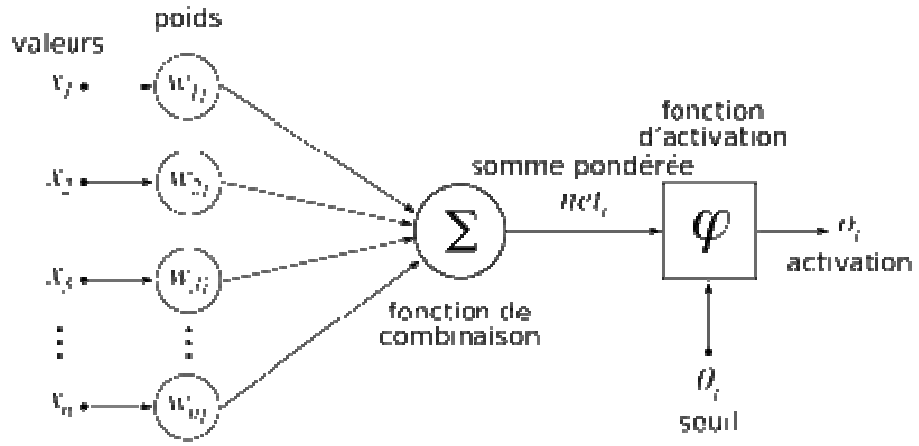

Fig.1. La structure d'un neurone artificiel (Haykin, S. 1994: Neural Networks)

Les entrées du réseau, notées $x_1, \ldots, x_n$ (se sont les variables explicatives du modèle), sont liées aux cellules de la couche suivante par des poids synaptiques w_{nj} déterminant l'effet du signal par le neurone *n* sur le neurone *j* ;

Chaque cellule de la première couche cachée (deuxième couche dans le réseau) va évaluer l'intensité de l'information qui lui parvient de la couche d'entrée en calculant la somme pondérée **:** $net_{j=} \sum w_{nj} x_n$

La cellule en question va traiter l'information puis elle va la transmettre aux cellules suivante à l'aide d'une fonction d'activation ζ qui anime le neurone en déterminant son activation ; et une activation o_j équivalente de la sortie de neurone. Elle est égale à: $o_j = \zeta (\sum w_{nj} x_n + \Theta_j)$

Où Θ_j se sont des paramètres à estimer généralement à l'aide d'une procédure d'apprentissage supervisé proposé par (Rumelhart et al 1986) connue sous le nom de l'algorithme de rétropropagation du gradient. Cet algorithme calcule l'erreur quadratique qui mesure l'erreur entre la sortie fournie par le réseau et la sortie désirée, mais cet algorithme introduit le nombre et la connectivité des unités cachées à priori, et détermine les poids par la minimisation d'un coût.

Le réseau obtenu est éventuellement élagué, ce qui en termes d'inférence non paramétrique qui permet de diminuer la variance (Le Cun et al., 1989, Geman et al., 1992, Hassibi al., 1993).

Nous constatons alors qu'avec une approche incrémentale on apprend au même temps le nombre d'unités et les poids, dans le cadre d'une architecture fixée, commençant généralement avec une seule unité.

Le perceptron multicouche (PMC)

Le perceptron multicouche (PMC) constitue le réseau le plus utilisé. Il s'agit d'un réseau de type (feedforward) composé de couches successives, ce type de réseau est très performant pour les problèmes de classification.

L'idée consiste à regrouper les neurones par couches interconnectées. Une première couche appelée couche d'entrée est composée d'un certain nombre de neurones dont la tâche est de recevoir l'information de l'extérieur. Ces informations sont transformées puis transmises aux neurones de la (ou des) couches intermédiaires qui vont effectuer certains traitement puis envoyer les résultats vers une dernière couche appelée couche de sortie.

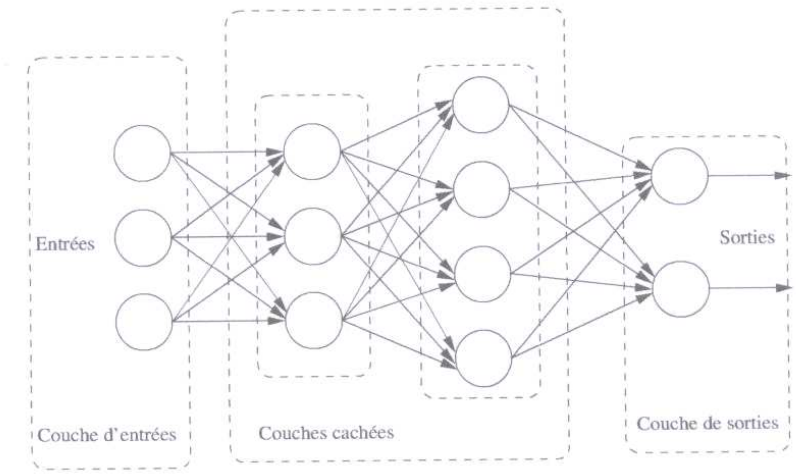

Fig. 2. Le perceptron multicouche (Borret. P 1991 : Réseaux de neurones artificiels)

Devant l'absence de règles théoriques permettant de déterminer l'architecture optimale d'un réseau de neurones pour une situation particulière, plusieurs auteurs (Mendelson 2001, Wierenga et Kluytmans 1994, Venugopal et Baets 1994, Shepard 1990) ont proposé des règles empiriques qui se basent sur un certain nombre d'essais obtenus en faisant varier le nombre et la taille des couches intermédiaires enfin, il est à noter que quelques tentatives de détermination automatique de la meilleure architecture ont été proposées.

MÉTHODOLOGIE

L'échantillon et les variables retenues

Notre étude porte sur 248 observations des femmes atteintes par la maladie du cancer du sein.

Les neurones d'entrées correspondent aux variables : épidémiologiques (age, origine, etat civil, contraception…), cliniques (signes inflammatoires, mamelon, gonglion…), para-cliniques (mammographie, métastase,…), classification TNM.

Quand la tumeur est suffisamment grosse, elle devient palpable. Parfois, en regard de la tumeur, la peau change d'aspect : ***peau d'orange, peau rétractée, ridée ou d'allure inflammatoire ; le mamelon peut aussi être déformé, être le siège d'un eczéma ou d'un écoulement de sang,*** avoir un aspect croûteux.

Il est aussi possible de découvrir des ***ganglions***, le plus souvent dans l'aisselle.

Des facteurs génétiques peuvent augmenter le risque (Antécédent familiaux)*:* on parle de gènes de prédisposition ou de risques de développer un cancer du sein. Plusieurs femmes dans une même famille (mère, grand-mère, tante, sœurs, filles) peuvent présenter un cancer du sein. La découverte d'un gène pathologique ne fait pas le diagnostic de la maladie, il indique seulement une élévation du risque de développer un cancer du sein.

Des risques hormonaux : il semble que les femmes ayant une puberté précoce, une **ménopause** tardive, n'ayant pas d'enfants ou ayant des enfants tardivement, n'ayant pas allaité, auraient un risque plus important d'avoir un cancer du sein.

La pilule contraceptive : Pour les contraceptifs oraux (CO), le risque serait un peu plus élevé *(+ 20 %)* pour les femmes ayant pris des CO pendant plus de 5 ans avant une première grossesse, de même il existe un sur-risque de cancer du sein chez les femmes utilisant un Traitement Hormonal de la Ménopause estroprogestatif. Ce sur-risque de cancer augmente avec la durée du traitement.

Antécédents personnels : ont plus de risques que les autres de présenter un autre cancer du sein dans leur vie (risque de deuxième cancer de *10 %).* Elles seront en général suivies attentivement après.

Les métastases : ce sont des cellules cancéreuses peuvent se détacher de la tumeur mère, aller dans le sang et se greffer sur des organes à distance et se multiplier pour leur propre compte.

Classification de la taille tumorale : T0, T1, T2, T3, T4, T4a, T4b, T4c, T4d, ***Nodule :*** N0, N1, N2, N2a, N2b.

La sortie correspond au type de cancer Maligne ou Bénigne.

Une couche cachée liant les entrées au sortie à travers une fonstion d'activation sigmoide :

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

Architecture du réseau

Cependant, comme pour toute application neuronale, un point crucial de la phase de modélisation reste la détermination de la structure du réseau. En effet, même si les travaux de (Cybenko, 1989 et Funahashi, 1989) ont montré qu'une seule couche cachée utilisant des fonctions d'activation de type sigmoïdale était suffisante pour pouvoir approximer toute fonction non linéaire avec précision voulue, rien n'est dit à priori sur le nombre de neurones cachées utilisé.

La problématique de la réduction du modèle a été initiée par (Zeigler, 1976) pour qui la complicité d'un modèle est relative au nombre d'éléments, de connexions et de calculs du modèle.

Selon Dunkin l'approche incrémentale permet de minimiser le nombre de nœuds à entrainer à chaque étape d'apprentissage, consiste à entrainer un réseau avec un neurone sur la couche en calculant l'erreur produite et le processus enchaine l'ajout des neurones tant que l'erreur au pas précédent est inférieure à un seuil donné et à chaque fois les poids du dernier neurone sont corrigés, le processus s'arrête ou l'incrémentation d'un nouveau neurone n'améliore plus la performance.

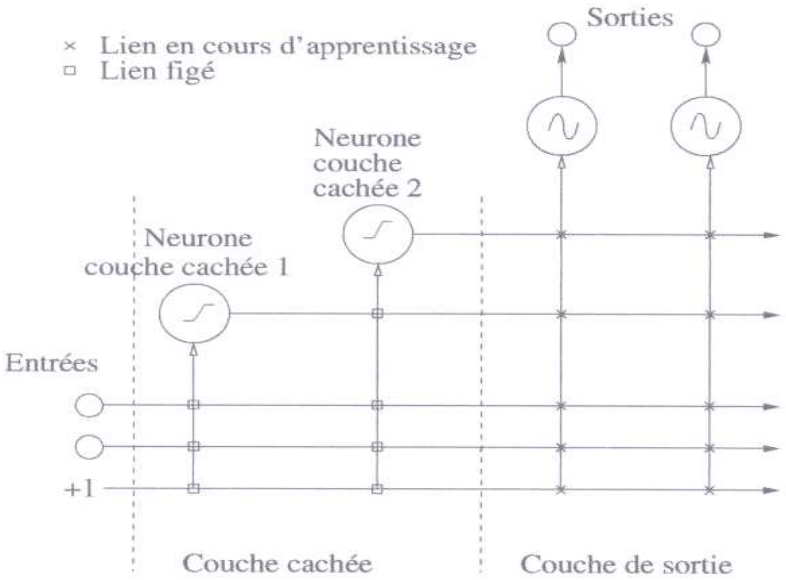

Fig.3. Algorithme incrémental (Dunkin, 1997)

Sélection des variables

Lorsqu'on conçoit un modèle à partir de mesures, il est indispensable que l'ensemble des variables d'entrée soit aussi réduit que possible.

Un grand nombre de techniques de sélection de variables a été proposé (Mc Quarrie et al, 1998).

La stratégie la plus naturelle, pour le choix d'un ensemble de descripteurs, consiste à partir d'un ensemble de variables candidates aussi grand que possible (c'est-à-dire le modèle complet) à comparer des performances de celui-ci à tous les modèles dont les entrées sont des sous ensembles de l'ensemble des variables candidates, et à choisir le meilleur selon un critère bien choisi.

On détermine généralement, par une analyse préalable de notre problème, les variables qui ont une influence sur notre phénomène étudié, on est alors condit à retenir un grand nombre de facteurs candidats, potentiellement pertinents, mais il est important de pouvoir

sélectionner ceux qu'ils sont réellement, on procède au classement, et on élimine toutes les variables qui sont moins bien classées (Stoppiglia, 2003).

Une fois les entrées sont classées par le procédé décrit on doit sélectionner, en fonction de ce classement, les variables qui doivent être conservées.

Il faut éviter de se tromper dans ce choix, car le fait de conserver des variables non pertinentes peut dégrader les performances du modèle, et le fait d'éliminer des variables pertinentes peut être tout aussi néfaste.

RÉSULTATS

Dans une première étape les 17 variables retenues ont été utilisées pour l'apprentissage d'un grand nombre de réseaux de type PMC à l'aide du logiciel ***Matlab*** ; ce logiciel comporte une fonction (*neural networks*).

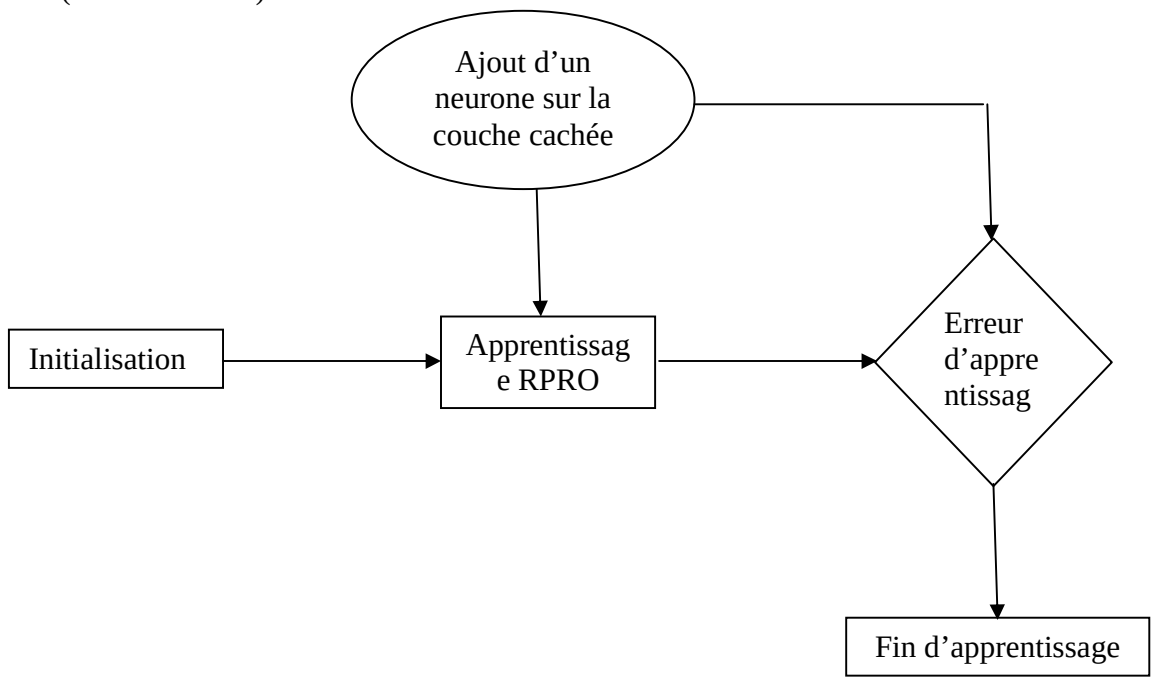

Fig. 4. Principe de l'algorithme d'apprentissage incrémental (Sauget, 2007)

Tableau 1. Le meilleur réseau utilisant 10 neurones cachés

Type	Inputs	Neurones cachés	Erreur	Performance : p*
PMC	17	10	0,00066944	0,98974359

Suite à la détection du meilleur réseau on va l'entrainer en suivant notre processus de sélection de variables (éliminer à chaque fois une variable) et comparer le nouveau résultat par rapport à ***la performance totale p*,*** puis classer les variables selon leurs degrés d'importance.

Tableau 2. Le classement des variables

Variables	Performance: pi	p*-pi	Rang
Antécédent personnel	0,292307692	0,69230769	1
Allaitement	0,323076923	0,66153846	2
Age 1er règle	0,338461538	0,64615385	3
Nodule	0,338461538	0,64615385	4
Mamelon	0,358974359	0,62564103	5
Antécédent familiaux	0,364102564	0,62051282	6
Ménopause	0,364102564	0,62051282	6
Contraception	0,394871795	0,58974359	7
Type de Peau	0,405128205	0,57948718	8
Métastase	0,425641026	0,55897436	9
Mammographie	0,430769231	0,55384615	10
Age 1ère Grossesse	0,435897436	0,54871795	11
Taille tumorale	0,461538462	0,52307692	12
Inflammation	0,471794872	0,51282051	13
Origine	0,528205128	0,45641026	14
Etat Civil	0,533333333	0,45128205	15

Notre architecture avec une seule couche cachée et 10 neurones cachés nous a permis de classer les facteurs de risques en deux parties plus la perte est importante plus la variable est importante comme : ***Antécédent personnel, Allaitement, Mamelon, Nodule, Age 1er règle (maturité après 15 ans), Antécédents familiaux, Ménopause, contraception, Type de Peau ;*** et celles qui génèrent une perte plus faible sont classées comme peu importantes : ***Métastase, Mammographie, Age 1ère Grossesse (âge supérieur à 35 ans), Taille tumorale, Inflammation, Origine, Etat Civil.***

CONCLUSION

Notre projet s'inscrit dans la branche d'intelligence artificielle en appliquant l'outil réseau de neurones comme une alternative intéressante aux statistiques traditionnelles dans l'aide au diagnostic médical plus particulièrement pour la maladie du cancer du sein.

Un premier travail a consisté à entrainer le réseau de neurones de type (*feedforward*) sur un échantillon de 248 patientes atteintes par la maladie du cancer du sein en Tunisie et 17 facteurs de risques en entrées, suite à un algorithme incrémental pour chercher le nombre optimal de neurones cachés qui s'est arrêté au niveau de 10 neurones cachés qui nous a offert un pourcentage de bon classement 98,97% , puis une deuxième étape c'est de sélectionner et classer les facteurs de risques selon leur effet direct sur cette maladie.

L'approche incrémentale au niveau de la sélection des variables nous permet de fournir un outil de prévision qui peut être utilisé par des professionnels du domaine médical.

Notre approche peut être améliorée dans le secteur de l'imagerie médicale, ainsi qu'une amélioration en appliquant d'autres algorithmes d'apprentissage comme ceux génétiques.

RÉFÉRENCES

Abdi, H. (1994). Les réseaux de neurones, Presses Universitaires de Grenoble, Grenoble

Armoni, A. (1998). 'Use the networks in medical diagnosis', Medical Diagnosis Computing 15:100-104.

Baxt (1995). 'Application of artificial neural networks to clinical medecine', The Lancet 346: 1135-1138.

Borret, P. et al, (1991). Réseaux de neurones artificiels : une approche connexionniste de l'intelligence artificielle, Teknea, Toulouse.

Cross, S., Harrison, R.F., Kennedy, R.L. (1995). 'Introduction to neural networks', The Lancet 346: 1075-1079.

Cybenko, G. (1989). 'Approximation by superposition of a sygmoidal function'. Math.control systems signals, 2(4): 303-314.

Davalo, E, Naim, P. (1990). Des réseaux de neurones, Eyrolles, Paris.

Drew, P., Monson, J. (2000). 'Artificial neural networks', Surgery 127:3-11.

Dunkin, N and Shawe-Taylor, J and Koiran, P (1997).A new incremental learning technique. In: Marinaro, M andTagliaferri, R, (eds.) Neural Nets: WIRN Vietri-96: Proceedings of the 8th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, 23-25 May 1996. (pp. 112 - 118). Springer: New York, US.

Diorio, C. (2005). Les facteurs de croissance analogues µa l'insuline, les apports en vitamine D et en calcium et la densité mammaire. Thèse de doctorat, Université Laval, Québec

Falkenberry, S., Legare, R. (2002). 'Risk factors for breast cancer'. Obstetrics and gynecology clinics of north America, 29(1): 159-172 (Allaitement maternale).

Funahashi, K (1989). 'On the approximate realisation of continous mapping by neural networks'. Neural networks, 2: 183-192.

Gairard' B., Mathelin, C., Schaffer, P. et al, (1998). 'Cancer du sein : épidémiologie, facteurs de risques, dépistage' Revue du Prat, 48 : 21-27.

Han, M., Snow, P. M., Partin, A.W. (2001). 'Evaluation of artificial neural networks for the prediction of pathologic stage in prostate carcinoma', Cancer 91(S8): 1661-1666.

Hassibi, B. Stork, D.G. Solla, S.A., (1993). Second order derivates for network puring: optimal brain surgeon. Advances in neural information precessing systems.

Haykin, S. (1994). Neural Networks. A Comprehensive Foundation. Macmillan, New York.

Le Cun, (1987). Modèles connexionnistes de l'apprentissage. PhD Thesis, Université Pierre et Marie Curie.

Sauget, M. (2007). Parallélisassions de problèmes d'apprentissage par des réseaux de neurones artificielles application en radiothérapie externe. Thèse de doctorat, université de Franche-Comté.

Maciej, A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, Georgia D. Tourassi (2009). 'Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance', Neural Network, 21(2-3): 427–436.

Rumelhurt, D.E. Hinton G.E., and Williams R.J (1986). Parrel disribeted processin, vol. 1-2. The MIT Press.

Stoppigli, Idan, Y., Dreyfus, G., (1997). Neural network aided portfolio management., in Industrial applications of neural networks ; World Scientific.

Geman,S., Bienenstock, E.,Doursat, R.(1992).'Neural networks and the bias/variance dilemma'. Neural Comp 4: 1-58.

Patil, S., Henry, J.W., Rubenfvie, M., Stein, P. D. (1993). ,Neural network in the clinical diagnosis of acute pulmonary embolism', Chest 10:1685-1689.

Thomas, P. and Bloch, G. (1998). Robust ptuning for multi layer perceptrons, IMACS/IEEE, CESA'98, Nabeul-hammamet.

Zeigler, B.P. (1976). Theory of modelling and simulation, Wiley, New York.