# A Review Of Machine Learning Techniques And Statistical Models In Anaemia

Jameela Ali, AbdulRahim Ahmad, Loay E. George, Chen Soong Der,Sherna Aziz

**Abstract:-** Blood diseases have in the recent past become a major cause of mortality and morbidity all over the world. Consequently, machine learning has emerged as one of the best and most fruitful methods of research in the present world, both in terms of proposing of new techniques with effective theoretical algorithms, and also in applying such methods in real life situations. From a technological view, it is evident that there are major changes in the world that occur at an ever increasing pace. This has seen the development of systems which can be easily adapt to the environment in an effective way by being practically applicable. These systems work through optimizing performance using certain algorithm in accordance with its maximization or minimization criteria, but also using experimental data instead of a given program. This study will identify the use of artificial neural networks, support vector machines and statistical models and methods in the recognition of iron deficiency that leads to anaemic conditions.

**Index Terms**: - Artificial neural networks, support vector machines, statistical models, Anaemic Red Blood Cell, cross validated

## 1 INTRODUCTION

The best feature about machine learning is that it combines knowledge form diverse fields such as pattern recognition involving; artificial neural networks, reinforcement learning, support vector machine, and decision tress, data mining, which entails modelling and time series prediction, statistics involving methods like Bayesian, Montecarlo and bootstrapping, or signal processing using Marvoc Method (Bornn & Tabet, 2010). This paper will look at the artificial neural networks together with the support vector machine method and how these methods in-cooperate the use of statistical methods (Adams, 2010). Due to the fact that, leucocytes and erythrocytes have a strong inter and intra variability in individuals, predicting of their ideal composition is the blood serum has proved to be a difficult task(QuagliniBarahona, & Andreassen, 2001). The aim of this study is to establish and analyze the impact of major parameters that are well known to influence erythrocytes and leucocytes composition together with iron presence in blood and to test the performance of the artificial neural network, support vector machines and statistical models in determination of anemic conditions in blood serum.

## 2ARTIFICIAL NEURAL NETWORKS

The performance of neural networks has been established as an effective strategy than the cox regression model in the prediction, determination and classification of clinical outcomes involving blood inflammations(Adams, 2010). Consequently, studies have shown that it is more accurate as compared to linear regression.


- Jameela Ali, E. George ,AbdulRahim Ahmad, Loay,Chen Soong Der,Sherna Aziz
- 1College of Graduate Studies
- 3,4College of Information Technology
- University Tenaga National -Malaysia
- 2,5Baghdad University-Iraq
- Jameela_ali65@yahoo.com, {abdrahim,chensoong}@uniten.edu.my, {loayedwar57@yahoo.com}


In this study, the optimal condition for a stable haemoglobin level has to be maintained between the range of 11 to 12 g/dl as recommended, and the concentration of the hemoglobin set above 12 g/dl. However, considering the possibility of thrombotic activities, it should not go above 14 g/dl (Quaglini Barahona, & Andreassen, 2001). The only way to determine the levels of these blood components is looking at aspects such as blood losses, dialysis efficiency, vitamin B12 deficiency, iron status, folic acid deficiency, pro-inflammatory cytokine activities, aluminum toxicity, and any previous treatments with angiotensin. Such parameters are essential in providing data used by the artificial neural networks (ANN) and support vector machines (SVM). In order to establish a non linear function that is continuous and expresses the interdependency of the data collected and erythrocytes and leucocytes levels, a series of neural networks are built, cross validated and trained using the Excel 4.32 software for neuro-solution, hence forming the artificial neural network (Suzuki, 2011). The ANN consists of one layer of in-put that collects in-put variables that should be predicted, one layer of output that collects the predictions which are known during training and unknown in validation and testing cases, and one or more layers that are hidden which perform the weight sum of the inputs, which then passes the results through the nonlinear function to reach the output layer (Adams, 2010). The individuals weights will be adapted progressively using the back propagation algorithm in order to reduce the big difference between the calculated values and the expected output. This means that the weights that provide an assurance of best results will be used to determine the performance of the ANN in establishing the levels of the leucocytes and erythrocytes (Suzuki, 2011). The purpose of applying ANN in hematology is to initiate the processes that human experts have developed to achieve reliable diagnosis that are separable from pattern analysis. This led to constructing of a hybrid system that combined rule based mechanisms and the ANN models in the recognition of microcytic anaemia (Bornn & Tabet, 2010). This involves the 3 layered program systems that use haematocrit, mean corpuscular volume, and coefficient variation of width cell distribution, as the inputs. The largely considered categories of anaemia include, iron deficiency (IDA), chronic disease anaemia (ACD), and hemoglobinopathy (HEM). These are particularly useful in the classification of

anaemia cases using the neural networks (Suzuki, 2011). The number of hidden units, momentum and learning rates are effectively varied to obtain a more appropriate classification. Information regarding cases of anaemia is then obtained and seventeen specific attributes are identified and used for the model training(Bornn & Tabet, 2010). Other anaemia raw data that is collected and preprocessed includes data cleansing, data preprocessing, and data selection. The validation data is used in monitoring the neuron network performance at the time of training, while test data is used in analyzing how the trained model is performing (Suzuki, 2011). An example of the range that can be used is input layers totalling to17 units, 15 units in the hidden layer, and 8 units for the output layer. The highest performed results were obtained when the hidden layer units were 15, 0.7 learning rate, and 0.1 momentum. Consequently, it emerges that there is 71.56% testing and 72.78% correctness. These show that the potential of multilayered perception in the recognition and predicting of anaemic cases and levels can be used by medical staff and haematologists (Suzuki, 2011).

## 3 SUPPORT VECTOR MACHINES

Patients suffering from chronic renal failure and complications, and are treated through maintaining their hemodialysis, commonly suffer from anaemia that is related to iron deficiency. In order to determine the iron status in such patients who are uremic, serum ferritin (SF) is considered to be the most effective indicator of the iron stored n the body. However, in some patients, even after being given iron therapy to manage their anemic conditions, they still do not improve in spite of being at low levels of SF (Suzuki, 2011). Other values that are specific to iron status such as; corpuscular indexes, saturation transfer, and serum iron are advocated for being reliable parameters in detecting iron deficiency. Support vector machine (SVM) is a recent techniques that in-cooperates the use of empirical data models that was developed by Vapnik. When this techniques is applied to specific medical problems, and classification or comparison of results is done there will be need to bring in statistical methods in such analysis (Quaglini Barahona, & Andreassen, 2001). The exclusion criteria used in the support vector machine include erythropoietin or therapy of androgen, presence of hepatic and inflammatory diseases and blood transfusion in recent months. The major classification was either suffering or not suffering from iron deficiency (Quaglini Barahona, & Andreassen, 2001). This is in accordance with the response of administering iron therapy, leading us to a dichotomous response identified by NR for No response meaning no iron deficiency and R for Response to iron deficiency. The variables used included serum ferritin, red cells (GR), hemoglobin (Hb), mean corpuscular volume (MCV), iron binding capacity, serum iron, and hematocrit. The problem with this classification was faced with the theory of the SVMs involving linear approach, instead of ensuring that the errors are minimized in the training data. The SVM conceptually implements the use of input vector which is non-linearly mapped onto a large dimensional feature space. This leads to a linear decision spaces. Special features in this space enable the SVM to embrace high generalization ability of training and learning of inflammatory conditions (Cortes, & Vapnik,1995). This is the basic feature that has elevated the SVM to be more effective than other data mine methods such as the classification trees. This high ability to generalize symptoms that have come into the machine in the form of data, use the polynomial input method of transformation. This exemplifies the SVM performance against the other classical learning algorithms that have been used in the recent past for clinical diagnosis (Cortes, & Vapnik, 1995).

For the linear classifier, the SRM principle has to be implemented using the OSH (optimal separating plane) and the hyper plane to minimize the distance between the two diverse classes. Only two points on the training set is critical in establishing the OSH (Bornn & Tabet, 2010). These two points are then referred o as the support vectors. From the numerical view, the SRM is reduced to find solutions to any constrained optimization problem (Quaglini Barahona, & Andreassen, 2001). The OSH varies in relation to other parameters such as the cost function regulation parameters and maximizes the margin while minimizing the number of points that are not in the range. Unpaired and multivariate relationship of the NR and R group showed major statistical differences for all the tested variables with the exception of hematocrit and hemoglobin. The SVM showed the highest value in terms of range sensibility (SE) (Bornn & Tabet, 2010).

## 4 STATISTICAL MODELS

The frequency and length of occurrence of iron deficiency cases are closely linked with morbidity in patients that are undergoing MD. Such rates are common outcomes in the statistical measurement and analysis of anemic conditions in patients. Consequently, more admissions of iron deficiency related symptoms in both CPD and MHD patients have been found to be involved with dialysis which may or may not be as a result of PEM or blood related chronic inflammation (Kopple, & Massry, 2004). Simple or multiple regression analysis is widely used in evaluating the measures of PEM rates in relation to hospitalization rates. The poisson regression model has also proved to be essential in the estimating hospitalization rates to a 95% confidence potential. Consequently, subgroups of anaemia or small sized sample cases can be predetermined using the vival-like statistical models. These models are mainly considered in monitoring the functional status and individual sense of well being which are generally referred to as the measurement of the quality of life (Kopple, & Massry, 2004). A number of statistical instruments are used in assessing the functional status and sense of well being and fall in a broad category of neural and serum inflammation detection instruments, while others are applied only to minimal extents. The Karnof sky score is a technique based on assessing the functional status of the cardiovascular system. It uses a simple questionnaire that consists of a short health care survey form divided into 36 questions (Zhang, & Rutgers the State University of New Jersey - New Brunswick. Graduate School - New Brunswick, 2008). The answers to these questions will then act as the input data to the statistical analysis in the models used. Other statistical instruments include the KDQOL-Kidney Disease Quality of Life instrument of survey which is mainly used for patients undergoing dialysis, and the Beck Depression inventory- BDI. Such instruments brings out results on the nature of the infection, stenosis or occlusion and the

frequency of the dialysis complications, which leads to establishing the survival rates of different dialysis modalities such as fistula, tunneled catheter, and grafts. These modalities may influence inflammatory and nutritional processes in patients that are undergoing MD. Consequently, the degrees of refractoriness of anemic conditions in patients that have undergone dialysis are other outcomes established by statistical models. The ESRD anaemia is considered as a multi-factorial disorder that is manage well through the recombination of iron therapy and erythropoietin (Zhang, & Rutgers the State University of New Jersey - New Brunswick. Graduate School - New Brunswick, 2008). Through statistical analysis it is easy to establish and classify the iron and EPO requirements that will maintain the optimum hemoglobin concentration from 11 to 12 g/L. From the above discussion, it is evident that apart from statistical methods being used in the diagnosis process in medical facilities, they are also used in the classification of certain inflammations and the kind of anemic condition. The surveys done above has elaborately established that statistical models in conjunction with the other machine learning instruments such as the SVM and ANN can be helpful to clinicians in the diagnosis and treatment of cardiovascular ailments and also in the learning process (Kopple, & Massry, 2004).

## 5 CONCLUSION

Statistical methods have proven to be useful in supporting medical diagnosis of diseases and inflammations in the presence of exhaustible and validated models of study. The models that have been discussed have confirmed that evaluation of all available indexes of iron status provides useful information in the diagnostic process of anaemic iron deficiency. The next step after making considerations that have been put forward in this discussion is to establish a multi-centre study that will have enlarged samples through which we will carry out an evaluation of the possibility of modifying the artificial neural networks and support vector machines to suite the study at hand. This will be essential in ensuring that the ordinal iron therapy response is taken into consideration and the data obtained taken through statistical analysis methods instead of using the dichotomy one. Owing to the limits of this study, we can conclusively suggest that support vector machines, and the artificial neural network through the use of statistical models have to be embraced as innovative ways in the current computer technology world as an effective means of approaching clinical problems such as anaemia recognition. However, the greatest limitations could be the complexity of the involved calculations when sampling large sizes, and in implementing relationships that are non linear. Consequently, this has been addressed through bring statistical models on board.

## References


[1]. Abedini, R., Zanganeh, I., & Mohagheghian, M. (2011). Simulation and Estimation of Vapor-Liquid Equilibrium for Asymmetric Binary Systems (CO2-Alcohols) Using Artificial Neural Network. Journal of Phase Equilibria and Diffusion, 32(2), 105-114

[2]. Abdolmaleki, P., Buadu, L. D., Murayama, S., Murakami, J., Hashiguchi, N., Yabuuchi, H., &

[3]. Masuda, K. (1997). Neural network analysis of breast cancer from MRI findings. Píloha diplomové práce, 15(5) 283-93.

[4]. Adams , D.C. (2010). Parallel evolution of character displacement driven by competitive selection in terrestrial salamanders. BMC Evolutionary Biology, 10(72), 1-10

[5]. Anagnostou, T., Remzi, M., Lykourinas, M., & Djavan, B. (2003). Artificial neural networks for decision-making in urologic oncology. Eur Urol, 43(6). 596-603

[6]. Atkov, Y. O., Gorokhova, G. S., & Sboev, G. A. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of cardiology, doi:10.1016/j.jjcc.2011.11.005

[7]. Baxt, W. G., Shofer, F. S,, Sites, F. D., & Hollander, J. E. (2002). A neural computational aid to the diagnosis of acute myocardial infarction. Ann Emerg Med, 39(4):366-73.

[8]. Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews, 11(10). 203-224

[9]. Benke, S. K., & Fallin, D. M. (2010) Methods: Genetic Epidemiology. Clinics in Laboratory Medicine, 30(4), 795-814

[10]. Bornn, L., Farrar, C.R., & Park, G. (2010). Damage Detection in Initially Nonlinear Systems. International Journal of Engineering Science, 48, 909-920

[11]. Bornn, L., & Tabet, A. (2010). Comment on "Particle Markov Chain Monte Carlo Methods". Journal of the Royal Statistical Society Series B, 72, 269-342

[12]. Cortes, C., & Vapnik, N. V. (1995). Support-Vector Networks. Journal Machine Learning, 20(3). Doi: 10.1023/A:1022627411411

[13]. El-Zammar, D., Yan, M., Huang, C., Fang, D., Petigara, F., Bornn, L., & Ngai, T. (2011)

Assessment and Management of Anaemia in a Population of Children Living in the Indian Himalayas: A Student-Led Initiative. UBC Medical Journal, 2(2), 12-18

[14]. Hannan, S. A., Bhagile, D. V., Manza, R. R., & Ramteke, J. R. (2010). Development of an Expert System for Diagnosis and appropriate Medical Prescription of Heart Disease Using SVM and RBF. International Journal of Computer science and Information security, 8(5). 122-201

[15]. Hong, D. W., Ji, F. Y., Wang, D., Chen, Z. T., & Zhu, H. Q. (2011). Use of artificial neural network to predict esophageal varices in patients with HBV related cirrhosis. Hepatitis monthly, 11(7). 544-547

[16]. Jorgensen, J. A., & Pirmohamed, M. (2011). Risk modeling strategies for pharmacogenetic studies. Pharmacogenomics 12(3), 397-410

[17]. Kenji Suzuki. (2011). Artificial Neural Networks-Methodological Advances and Biomedical Applications. Introduction to the Artificial Neural Networks. Retrieved from <http://www.ltfe.org/wp-content/uploads/2011/04/Artificial_Neural_Networks_-_Methodological_Advances_and_Biomedical_Applications.pdf>

[18]. Kopple, D. J., & Massry, G. S. (2004). Kopple and Massry's nutritional management of renal disease. Philadelphia: Lippincott Williams & Wilkins. Print.

[19]. Lisboa, P. J. G. (2004). Neural networks in medical journals: current trends and implications for BioPattern. Proc. 1st European Workshop on Assessment of Diagnostic Performance (EWADP), 7(9). 99-112

[20]. Magalhães, R. J. S., & Clements, A. C. A. (2011). Mapping the Risk of Anaemia in Preschool-Age Children: The Contribution of Malnutrition, Malaria, and Helminthes Infections in West Africa. PLoS Med, 8(6) doi:10.1371/journal.pmed.1000438

[21]. Mandal, I., & Sairam, N. (2012). Accurate Prediction of Coronary Artery Disease Using Reliable Diagnosis System. Journal of Medical Systems, 2(3). 93-235 DOI: 10.1007/s10916-012-9828-0

[22]. Min, L., Wu, W., Joseph, R., Fulton, D.B., Berg, L., & Andreotti, A.H. (2010). Disrupting the intermolecular self-association of Itk enhances T cell signaling. Journal of Immunology, 184, 4228-4235

[23]. Nasr, B. M., & Chtourou, M. (2010).Training recurrent neural networks using a hybrid algorithm. Neural Computing & Applications. 21(1), 1-203

[24]. Quaglini, S., Barahona, P., & Andreassen, S. (2001). Artificial intelligence in medicine: 8th Conference on Artificial Intelligence in Medicine in Europe. Berlin: Springer. Print.

[25]. Ranganath, H., & Gunasekaran, N. (2004). "Artificial Neural Network Approach in Estimation of Hemoglobin in Human Blood." International Computer Engineering Conference on New Technologies for the Information Society, ICENCO, Cairo, 341-344

[26]. Shi, H., Lu, Y., & Du, J. (2011). Application of Back Propagation Artificial Neural Network on Genetic Variants in Adiponectin ADIPOQ, Peroxisome Proliferator-Activated Receptor-γ, and Retinoid X Receptor-α Genes and Type 2 Diabetes Risk in a Chinese Han Population. Diabetes technology & therapeutic., doi:10.1089/dia.2011.0071

[27]. Shukla, A., Tiwari, R., & Kala, R. (2010). Towards Hybrid and Adaptive Computing. Studies in Computational Intelligence, 307, 31-58

[28]. Tahmasebi, P., & Hezarkhani, A. (2011). Application of a Modular Feedforward Neural Network for Grade Estimation. Natural Resources Research. 20(1). 25-32

[29]. Tong, L. L. D., & Schierz, C. A. (2011). Hybrid genetic algorithm-neural network: feature extraction for unpreprocessed microarray data. Artificial intelligence in medicine, 53(1). 47-56 doi:10.1016/j.artmed.2011.06.008

[30]. Wehe, Chang, W.-C., Eulenstein, O., & Aluru, S. (2010). A scalable parallelization of the gene duplication problem. Journal of Parallel and Distributed Computing, 70, 237-244.

[31]. Xu, M., Brar, H., Grosic, S., Palmer, R., & Bhattacharyya, M.K. (2010). Excision of an active CACTA-like transposable element from DFR2 led to variegated flowers in soybean. Genetics: 184, 53-63.

[32]. Zahir, S., Rejaul, G. C. & Payne, W. (2006). "Automated Assessment of Erythrocyte Disorders Using Artificial Neural Network". IEEE International Symposium on Signal Processing and Information Technology, Vancouver, 776-780.

[33]. Zhang, X., Wang, Z., Tang, L., Sun, Y., Cao, K., & Gao, Y. (2011). Support vector machine model for diagnosis of lymph node metastasis in gastric cancer with multidetector computed tomography: a

preliminary study. BMC cancer. Retrieved from <http://www.biomedcentral.com/1471-2407/11/10>

[34]. Zhang, G. P. (2000). Neural Networks for Classification: A Survey. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications And Reviews, 30(4). 451-462

[35]. Zhang, Q., & Rutgers The State University of New Jersey - New Brunswick. Graduate School - New Brunswick. (2008).Maternal anaemia and adverse pregnancy outcomes. Michigan: ProQuest