

The Artificial Neural Networks for Cancer Research in Prediction & Survival (ANNCRIPS)

Karan Kamdar*, Amit Mathapati**

* Vivekanand Institute of Technology, Mumbai, India

** Vivekanand Institute of Technology, Mumbai, India
karan@stople.com ; amitmathapati@yahoo.com

Abstract— Prostate cancer is the most common cancer among men in age 50 or older, excluding skin cancer. Current methods of screening for prostate cancer carried out through blood tests & presence of high PSA lead to a high percentage of false positive test result (FPTRs) which can be reduced by employing intelligent Artificial Neural Networks. The goal of our research paper and the parallel undertaking of its practical implementation is to develop a mathematical model to improve prostate cancer detection and staging systems and finally to present a deploy ready marketable solution based on the model which can be installed across various screening centers, hospitals and research organizations.

to be accompanied with a digital rectal exam where a biopsy of the prostate may be required or some other method to confirm the presence of cancer. These methods work in situations where there is strong verifiable evidence of the presence of cancer. However in extreme cases where there is no clear signal of cancer occurrence a high number of false positive test results (FPTRs) may be produced resulting in repeated screening and diagnostic procedures. Hence, the goal of this paper is to employ Intelligent Artificial Neural Networks to build a deploy ready marketable mathematical model to reduce the non-required trial and error methods and expedite early cancer diagnosis and treatment

I. INTRODUCTION

A. The Genesis - Why is the emphasis on prostate cancer ?

Prostate cancer is the most common cancer among men, excluding skin cancer.

American Cancer Society (ACS) estimates for 2005 include 232,090 new cases of prostate cancer in the U.S.

Year 2005 estimates include 30,350 deaths occurring from prostate cancer in the US alone, making it the second leading cause of cancer death in men.

All men are at risk for prostate cancer. The risk increases with age, and family history also increases the risk.

African-American men have about a 70 percent higher incidence rate of prostate cancer than Caucasian men, and nearly a two-fold higher mortality rate than Caucasian men.

B. The Need for efficient prostate cancer screening and diagnostic methods

Increasingly, physicians are relying on the results from the PSA (prostate specific antigen) blood test to diagnose prostate cancer. When the prostate gland has a hint of cancer-producing cells, a detectable substance, namely PSA, is produced in the blood. An abnormally high PSA level alerts the physician to the possible presence of prostate cancer. Although the PSA blood test has been refined and developed, the physician cannot truly rely on the results of this test alone and so the PSA blood test has

II. THE ANNCRIPS

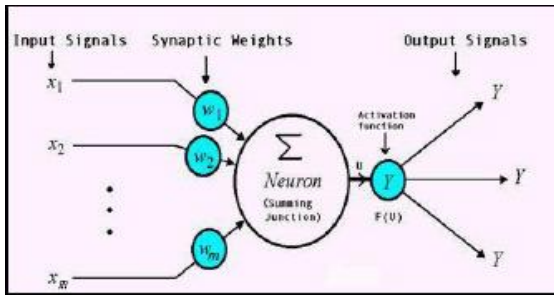
A. Project Overview

The Artificial Neural Networks for Cancer Research in Prediction and Survival (ANNCRIPS) is a voluntary effort started by the authors of this paper. The idea is to build a mathematical model to improve prostate cancer detection and staging systems. Here the basis is to add intelligence to Artificial Neural Networks and produce a deploy-ready mathematical model that revolves around the concept of ANNs. The actual implementation of the model would require building a standalone software application which would be installed across various screening centers, hospitals and research institutions. The project is in its intermediate stage and the team is working on the development and implementation stages to bring the product to the masses. C is the choice of programming language to be used for the reasons of its strong mathematical library and Matlab is used as the simulation software package to test our software runs. This research project is a long term initiative involving successive refinements to the first version of our model.

B. Artificial Neural Networks



The Artificial Neural Networks (ANNs) are information processing paradigms inspired by the way biological nervous systems such as the brain process information. The figure 1.1 shows such a typical example of the human brain's neural network. ANNs adopt this interconnected neuron network to perform complex computations. The typical ANN model is as shown below:



$$u = \sum_{j=1}^m w_j x_j$$

$$Y = F(u)$$

Fig 1.2

C. Overview of Prostate Cancer

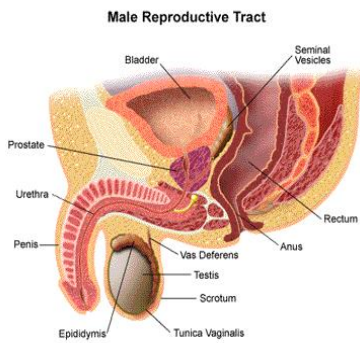


Fig 1.3

The prostate is a sex gland in men. It is about the size of a walnut, and surrounds the neck of the bladder and urethra - the tube that carries urine from the bladder. It is partly muscular and partly glandular, with ducts opening into the prostatic portion of the urethra. The prostate gland secretes a slightly alkaline fluid that forms part of the seminal fluid, a fluid that carries sperm.

Prostate cancer is the cancer beginning in the prostate which may remain in the prostate gland or spread to other internal organs. Screening & Diagnosis are carried out through blood tests measuring the level of PSA and alternatively accompanied by a DRE or digital rectum exam. After initial diagnosis further staging system such as TNM may be undertaken to pinpoint the location of cancer and the degree of its spread into other regions.

D. Linking ANNs to Prostate Cancer Analysis

The current method for prostate cancer detection is finding the percentage of tPSA (total prostate specific antigen). However there are some core drawbacks related to this approach. The first being that only cancer confined to the prostate gland can be detected and the second being that additional information such as the degree of cancer spread to other internal organs requires more tissue samplings.

A possible solution to these problems may be found in finding the percentage of fPSA (free prostate specific antigen) to suggest the degree of spread of prostate cancer to other parts. However both the above solutions have the following drawbacks:

- They cannot provide specific diagnostic results on a person to person basis
- There are unable to learn from existing diagnostic patterns and associate them with new input data from different patients falling under the same category based on age group, degree of cancer spread, number of years till cancer was diagnosed and other generic parameters.
- Both the systems provide accuracy levels ranging from satisfactory to moderately efficient and hence repeated diagnostic measures to overcome the occurrence of false positive test results have to be undertaken.

E. Building the ANN model

Our ANN Model is based on the concept on Multilayer Perceptron (MLP). It consists of a network of processing elements or nodes arranged in layers.

Principle - Input pattern presented at the input layer causes network nodes to perform calculations in the successive layers until an output value is computed at each of the output nodes from which the most significant is selected

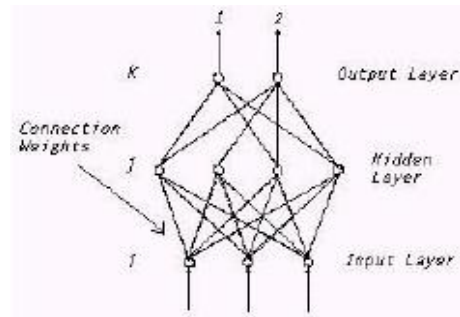


Fig 1.4

- Working

$$input_j = \sum_i w_{ji} out_i$$

- Input to node j :

$$out_j = f(input_j)$$

- Output of node j :

- This continues through all the layers of the network until output layer is reached and output vector is computed. The function f denotes

activation function of each node. A sigmoid function is generally used

$$f(x) = \frac{1}{1 + \exp(-x)}$$

F. Extending the MLP methodology

Our neural network model used is a multilayer perceptron (MLP) network composed of one input layer with four primary preprocessed variables (tPSA, fPSA, prostate volume, DRE).

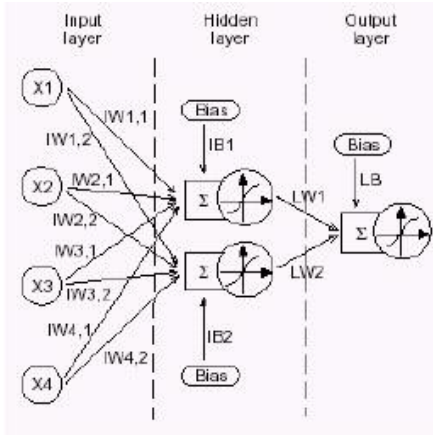


Fig 1.5

There is one hidden layer with two neurons, and one output layer with one neuron giving the output value that is a measure of the probability of cancer.

The 13 parameters (weights) associated with each input layer are to be optimized and the **Activation function** used in the hidden and output layers is hyperbolic tangent sigmoid function.

$$[a = \tanh(s) = (e^s - e^{-s}) / (e^s + e^{-s})]$$

Hence the output values are between -1 to 1

The formula for whole network thus becomes :

$$a1 = \tanh (IW1, 1 * X1 + IW2, 1 * X2 + IW3, 1 * X3 + IW4, 1 * X4 + IB1)$$

$$a2 = \tanh (IW1, 2 * X1 + IW2, 2 * X2 + IW3, 2 * X3 + IW4, 2 * X4 + IB2)$$

$$aout = \tanh (LW1 * a1 + LW2 * a2 + LB)$$

After model build-up, the respective weights are initialized and the inputs from the patient data set are entered. Inputs are actual data of patients suffering from prostate cancer. The whole network is now trained using Levenberg-Marquardt & Bayesian optimization techniques to recognize and associate patterns of input with desired outputs indicating the correct cancer risk.

After this the model is run with internal processing in the network taking place and resultant output taking the value between 0 (low PCa risk) and 1 (high PCa risk). In some cases, the value is <0 or >1 which is not relevant. Multiple runs of the model are now executed and a statistical curve indicating the values of PCa risk is plotted. If the curve indicates high occurrence of values closer to 1 then existence of prostate cancer is confirmed. The resultant data is not excreted but is feedback into the system so that a higher accuracy of the statistical curve applicable to patients belonging to similar data set is made possible.

Intelligent learning and network training is the key to our ANN model's success. If this is not done then undesirable results may be obtained. E.g. if the network has been trained only once for a person having high PSA level due to non-cancerous cells then it wont indicate any cancer risk for persons falling into the same characteristic group and actually having a high PSA level. Hence repeated training and learning should be employed widely so that the model learns to associate and de-associate itself at the same time from different pattern sets. The model can also be categorized as stochastic for the reasons of having input data set which is variable across different patients who are tested.

G. Training Procedure

The following training procedure is used for our MLP based ANN model

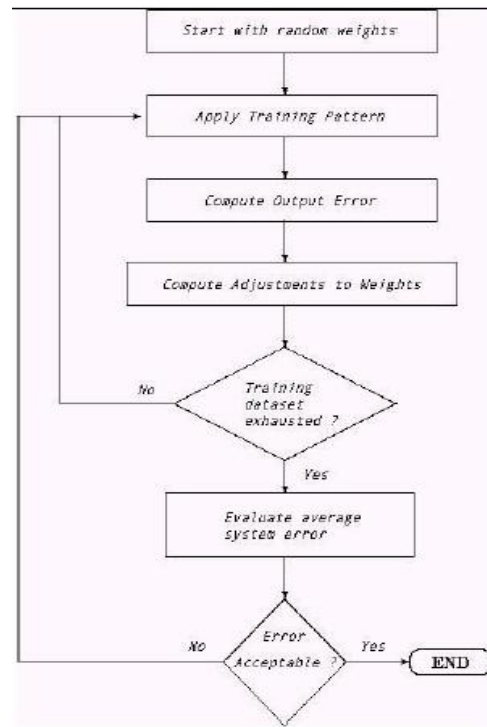


Fig 1.6

H. 5 Steps to implement the ANNCRIPS model

■ Step 1 – Obtaining the Input Data Set

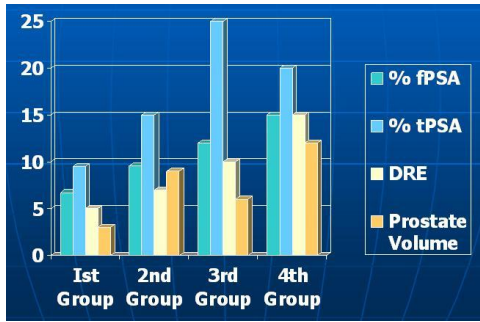


Fig 1.7

■ Step 2 – Adjusting Initial Weights

- i) The model is fitted on 75% and tested on 25% patients in each training set.
- ii) Randomization, fitting and testing sessions repeated 5 – 6 times and the weights producing smallest sum of squared errors on the initial test set are selected as initial weights for the final training of the MLP.

■ Step 3 – Train the Network

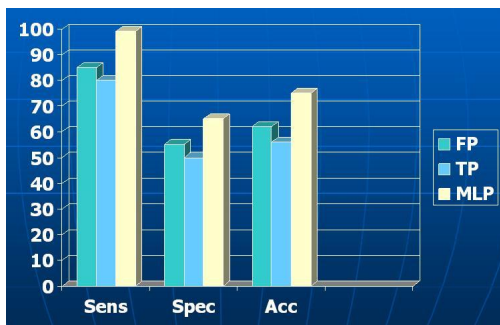
- Actual Training of Network begins with estimation & validation samples.

■ Step 4 – Test the Network

If validation objective is acquired, network trained again till best performance parameter values are obtained.

■ Step 5 – Validating & Generalizing Network Efficiency

I. Results of our ANN on a comparative basis



- **Sensitivity** : Ability of the model to detect prostate cancer early.
- **Specificity** : Efficiency of avoiding repeated tissue samplings.
- **Accuracy** : Proportion of subjects with a correct test result.

From the above graphical chart it is seen clearly that our ANN model would better the fPSA or tPSA models quite substantially.

- **ROC Curve** : Our ANN model also occupies a greater area under curve as compared to fPSA & tPSA models

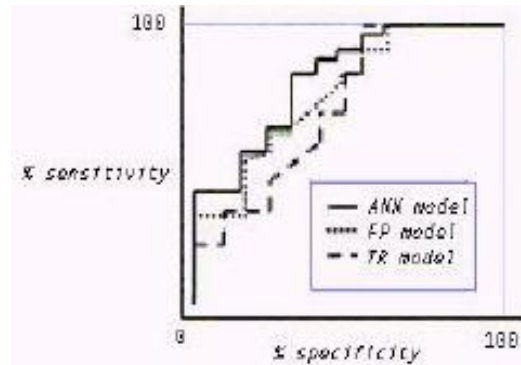


Fig 1.8

III. THE PROPOSED ANNCRIPS PRODUCT

The following is a sample of how our product would look like once it has been implemented. The team is currently working on producing a software solution by mapping the proposed mathematical model to the C language and building simulation runs through Matlab to exploit the core mathematical functionality of the two programming packages. We also find tremendous scope in terms of improvement in the mathematical model, training procedure, testing and validation steps.

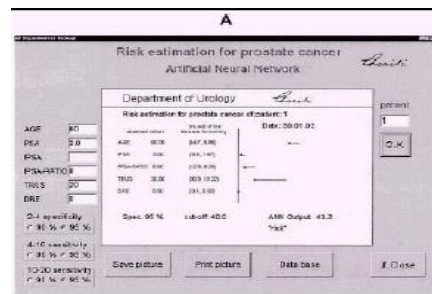


Fig 1.9

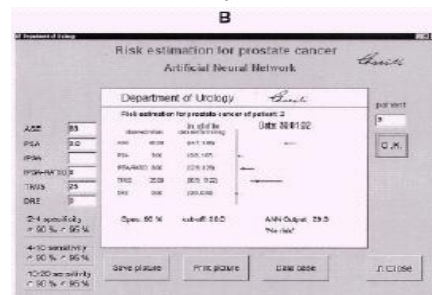


Fig 2.0

IV. CONCLUSION

- Prostate cancer is the most common form of cancer among men according to statistical study results.
- Screening is a very rough estimate for cancer risk
- Further staging systems such as A, B, C, D, TNM lack required efficiency.
- Solution: Our proposed ANN model is far more efficient in predicting prostate cancer risk and reducing the no. of false positive test results.
- Our model is developed to aid or substitute the current diagnosis and prognosis methods.
- Our software test runs have shown that highly accurate ANNs based on extending the MLP model can detect prostate cancer early & reduce unnecessary tissue samplings as compared to current methods such as fPSA and tPSA.

Influencing factors:

- Larger no. of input variables from the patient data set can be supported.
- Interconnecting relationships between these input variables and thereby formation of re-usable patterns can be established.
- The ability of the neural network to be trained time and again can be exploited thereby increasing accuracy each time.
- A deploy-ready easy to install software solution would be made available in the form of our ANNCRIPS product which can be installed across various screening centers, hospitals and research institutions.

V. ACKNOWLEDGEMENT

The team would like to thank Prof. Balakrishnan who is the Head of Computer Science Department at the Vivekanand Institute of Technology, Mumbai for his guidance and support when we initially looked for references and sources of information on this research project. We look forward to his guidance in the years to come. We would also like to thank the Student Library at the Vivekanand Institute of Technology, Mumbai for providing exhaustive resources for Prostate Cancer Research and Neural Network references.

REFERENCES

- [1] For Research on Neural Networks and Prostate Cancer : Vivekanand Institute of Technology's Student Library, <http://www.vesit.edu>
- [2] For Statistics on prostate cancer : <http://www.health.uab.edu/show.asp?durki=15376>.

R