Prediction and Detection of Rheumatoid Arthritis SNPs Using Neural Networks

K.B Ramesh[1], S.J Shibani Prasad[2], Dr. B.P. Mallikarjunaswamy[3], Dr. E.T.Puttaiah[4]

1. Associate Professor, Dept. Of Instrumentation Technology, RVCE, Bangalore, India.
2. IV sem M.Tech, Dept. Of Instrumentation Technology, RVCE, Bangalore,India.
3. Professor, Dept. of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumkur, India.
4. Professor, Dept. of Environmental Science, Vice Chancellor, Gulbarga University, India.

Abstract: *Rheumatic Arthritis (RA) is the most common disease found in the majority of the populations next to diabetes. RA is a chronic systemic inflammatory disease that primarily affects the synovial joints. The genes do contribute for the development of RA and it varies among individuals and between populations in different age group. It is necessary to know the gene factors that are associated with the disease for the better understanding of the underlying causes of the disease. Since RA is an auto-immune disease, it is important to identify the responsible SNPs of RA in order to detect and predict the disease well in advance. Prediction of RA helps in the early diagnosis of the disease and helps in improving the quality of life. This paper gives a detailed review of the existing methods that are used in the prediction of RA SNPs and also an ideology which works on the concept of Neural Network to detect and predict RA if a DNA sequence is given. The outcome of this would help doctors, genetic scientists, pharmacists in understanding the characteristic gene responsible for RA and provide proper diagnosis method and in discovering new drugs.*
Keywords: Rheumatoid Arthritis, SNPs, Neural Networks.

I.   Introduction

Diabetes is the most common metabolic disorder affecting populations in all geographical regions of the world. The prevalence of diabetes is influenced by genetic, ethnic and socioeconomic factors. The World Health Organization has projected that the prevalence of diabetes is increasing in epidemic proportions especially in developing countries. India has the highest number of people with diabetes in the world. On the other hand Rheumatoid Arthritis is the next most common condition affecting 1% of Indian population. The dark part of Rheumatoid Arthritis lies in the fact that doctors are not aware of what exactly causes the disease and it is said to be witnessed at any age from children to elders. The most common age groups prone to Rheumatoid Arthritis are between 30 and 50. There is no single diagnostic clinical or laboratory test to confirm the diagnosis of Rheumatoid Arthritis. Hence the prediction of Rheumatoid Arthritis has been a challenge since long time and techniques for the early detection of the negative Rheumatoid factor is being initiated and several studies have been made on the same. Also the research suggests that there is a connection between Rheumatoid Arthritis and Diabetes. But the nature of that connection is still unclear. Although the exact cause of RA is unknown, some possible factors including infections, viruses, diet and possible injury have not received scientific backing. Among the men mostly affected are smokers. Smoking is undoubtedly an important factor in causing RA. Although rheumatic arthritis has a genetic component, possibility of passing on to a child is extremely low. The precise quantification of cartilage damage and bone destruction (i.e., joint space narrowing and erosions) caused by RA is a decisive factor in treatment, whether with aggressive treatment strategies (e.g., methotrexate) or biologic agents, such as anti- Tumor Necrosis Factor alpha (anti-TNF alpha) drugs. The diagnosis of early RA is very difficult. Furthermore, it is of great importance to compare the anti-inflammatory effects of different agents, which retard radiographic progression, in clinical multicentre whereas ultrasound and magnetic resonance imaging are used to monitor synovitis and short-term changes, radiography is used as the standard method to monitor the long-term progression of RA. Hence a method to detect and predict the presence of RA at its early stage is necessary for proper diagnosis.

II.   Methods

Discrete Karhunen–Loeve transform (DKLT):

A generalized classification methodology is developed to predict the presence or absence of a multifactorial disease from a set of risk factors thought to be correlated with the disease. The methodology includes fusion to combine risk factors into a single feature vector, normalization to overcome the problems associated with fusing features which have different formats and ranges, discrete Karhunen–Loeve transform

(DKLT)-based transformation[1] to facilitate parametric classifier development, the selection of features with high interclass separations, and the design of parametric classifiers. The validity of the method is demonstrated by applying it to predict the occurrence of gout from 14 risk factors.

Deciding whether a test patient has or does not have the disease can be formulated as a hypothesis testing problem in which:

H0 : Vt = V n; the patient does not have the disease;
H1 : Vt = V m; the patient has the disease.

In order to facilitate classifier development, the normalized patient vector which may have both real and binary risk factors, can be transformed so that the transformed vector contains real-valued elements. As a result, it would be much easier to make meaningful assumptions for the conditional densities of the patient vectors under the two hypotheses. The DKLT is a suitable transformation because each element in the transformed vector is a linear combination of the risk factors in the patient vector. Let ~ V n and ~ V m be the transformed patient vectors. That is

~ V n =_V n
~ V m =_V m

where _ is the (s _ s) generalized DKLT transformation matrix consisting of the eigenvectors of the covariance matrix of the mixture of the two classes. The classification accuracy is estimated as the fraction of the number of correctly classified vectors in the test set. The random partitioning can be repeated H times (trials) and the classification accuracy is then estimated by averaging the resulting H classification accuracies.

The classification accuracy is given by,

$$\alpha_H = \left[\left(\frac{1}{H}\right)\sum_{h=1}^{H}\alpha_h\right] \times 100\%$$

where

$$\alpha_h = \frac{\text{number of correctly classified vectors in the test set}}{\text{number of vectors in the test set}}$$

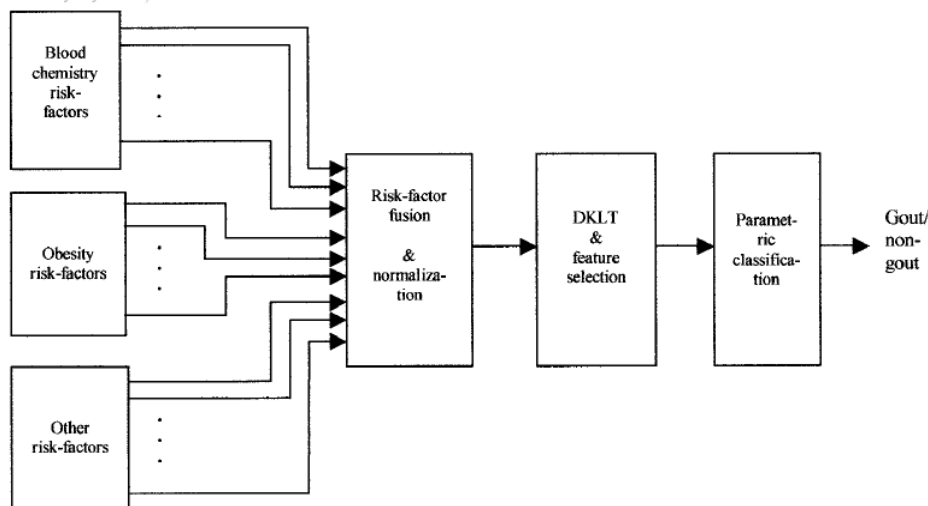in trial h, $h = 1, 2, \ldots, H$.

Fig.1. Block diagram of the risk-factor fusion methodology to predict gout.

In summary, it can be concluded that the generalized risk-factor-fusion-based classification approach developed by this method is quite effective in predicting the occurrence of a multifactorial disease. Because it is difficult to clinically diagnose such diseases accurately, the classification methodology can be used by a physician to support or assist him/her in diagnosing a multifactorial disease.

a) **Disease-disease relationships for rheumatic diseases:**

The MEDLINE database (Medical Literature Analysis and Retrieval System Online) contains an enormously increasing volume of biomedical articles [2]. There is urgent need for techniques which enable the discovery, the extraction, the integration and the use of hidden knowledge in those articles. Text mining aims at developing technologies to help cope with the interpretation of these large volumes of publications. Co-occurrence analysis is a technique applied in text mining and the methodologies and statistical models are used

to evaluate the significance of the relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications. The results reveal hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE, thereby exposing relationships that can provide important additional information for medical experts and researchers for medical decision-making.

b) Evolvable Rough-Block-Based Neural Network and Its Biomedical Application to Hypoglycemia Detection System :

This method focuses on the hybridization technology using rough sets concepts and neural computing for decision and classification purposes. Based on the rough set properties, the lower region and boundary region are defined to partition the input signal to a consistent (predictable) part and an inconsistent (random) part. In this way, the neural network is designed to deal only with the boundary region, which mainly consists of an inconsistent part of applied input signal causing inaccurate modeling of the data set. Owing to different characteristics of neural network (NN) applications, the same structure of conventional NN might not give the optimal solution. Based on the knowledge of application in this paper, a block-based neural network (BBNN) is selected as a suitable classifier due to its ability to evolve internal structures and adaptability in dynamic environments[3]. This architecture will systematically incorporate the characteristics of application to the structure of hybrid rough-block-based neural network (R-BBNN).

A global training algorithm, hybrid particle swarm optimization with wavelet mutation is introduced for parameter optimization of proposed R-BBNN. The performance of the proposed R-BBNN algorithm was evaluated by an application to the field of medical diagnosis using real hypoglycemia episodes in patients with Type 1 diabetes mellitus. The performance of the proposed hybrid system has been compared with some of the existing neural networks. The comparison results indicated that the proposed method has improved classification performance and results in early convergence of the network.
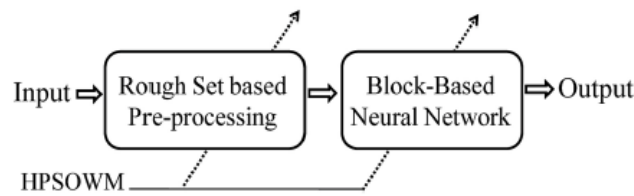

Fig 2: Proposed rough-block-based neural network (R-BBNN).

The rough sets theory predominantly deals with the classification analysis of imprecise, inconsistent, or incomplete information. It offers effective methods that are applicable in many branches of artificial intelligence (AI) technologies. Its soundness and usefulness has been proven in many real life applications.

BBNN is a network design that is more flexible for changing the structure depending on the signal flow between blocks. It can be represented by a 2-D array of blocks. Each individual neuron block works as a basic signal processing unit that is composed of a feed forward neural network having four variable input/output nodes. Owing to different characteristics of neural networks (NNs), a traditional neural network with a common structure may not be able to handle every application. Based on the knowledge of application, BBNN is selected as a suitable classifier due to its modular characteristics and ability in evolving the size of the network by adding more basic blocks. The structure of the network simply corresponds to determining signal flows between each block. In addition, conventional NNs normally offer less DOF, i.e, the network structure may not be evolving during the training process. Conceptually, the number of degrees of freedom allows the network to be more adaptive with the characteristics of application. If less degrees of freedom, the model becomes less accurate. But for BBNN, their weight and structure are evolving along the training process and offer more DOF. Consequently, the model becomes more adaptable with the characteristic of application and gives a more accurate result. Easy implementation of reconfigurable digital hardware platform with field programmable logic arrays (FPGAs) is another attraction of evolved BBNN. In conventional neural network, the implementation in digital hardware is not an easy task due to nonlinear activation functions and real-valued connection weights. The selection on the number of hidden layers is the same as the way which is chosen for conventional feed forward neural network. In a conventional neural network, the number of hidden nodes is chosen by a trial and error method until the optimal solution is obtained. The same conditions are applied, so an optimal number of hidden nodes is selected by a trial and error process till the desired condition is met. The network parameters are weights, biases, and transfer function parameters are *wij* , *bj* , *α* and *β*. A search method hybrid particle swarm

optimization with wavelet mutation (HPSOWM) is used to find optimal values. The position $x_{pj}(t)$ and velocity

$$
\begin{aligned}
x_j^p(t) &= x_j^p(t-1) + v_j^p(t) \\
v_j^p(t) &= k.\left(m.v_j^p(t-1) + \varphi_1 r_1\right).\left(\tilde{x}_j - x_j^p(t-1)\right) \\
&\quad + \varphi_2 r_2 \left(\hat{x}_j - x_j^p(t-1)\right)
\end{aligned}
$$

$v_{pj}(t)$ of HPSOWM are given as follows:
This architecture has been designed in such a way that the rough set method can greatly accelerate the network training time and improve its prediction accuracy.

III.   Future Scope

From the above reviewed techniques, it is clear that detection and classification of the SNPs can also be done by implementing the above methods, except for that it should provide an accurate result based on which the diagnosis of RA is carried out. Identifying the SNPs responsible for RA can be done by considering a system which implements the concept of artificial Neural Networks in which a false set of cancer is taken as a reference in contrast to RA. A database is created which stores the SNPs related to cancer and RA. Using this database, required features are extracted which paves way in understanding the common genomic sequence that are responsible for both the diseases and thus helps in the prediction of RA if a DNA sequence is given. This system facilitates the pharmacists in discovering new drugs for RA and thus improving the quality of life.

IV.   Conclusion.

Diagnosing RA in its early stage can be difficult. There is no precise test that can clearly identify RA; hence a system that helps in the early prediction and detection of RA is necessary. This innovative approach is based on the merit of roughset and BBNN, BBNN is selected as a suitable classifier due to its ability to perform simultaneous optimization of both structures and weights. To enhance the performance of the proposed R-BBNN4 system, a HPSOWM operation can used to optimize the design parameters. Based on this architecture a tool for the prediction and detection of Rheumatoid arthritis can be developed which thereby would serve the biomedical field research in accomplishing other disease related activities as well.

References

[1].   James Phegley, Kyle Perkins, Lalit Gupta and J. Kevin Dorsey," Risk-Factor Fusion for Predicting Multifactorial Diseases", IEEE Transactions On Biomedical Engineering, Vol. 49, No. 1, January 2002.
[2].   Andreas Holzinger , Klaus-Martin Simonic, Pinar Yildirim," Disease-disease relationships for rheumatic diseases", Web-based biomedical textmining and knowledge discovery to assist medical decision making, IEEE 2012.
[3].   Phyo Phyo San, Sai Ho Ling, Nuryani, Hung Nguyen," Evolvable Rough-Block-Based Neural Network and Its Biomedical Application to Hypoglycemia Detection System",IEEE journal 2013.
[4].   Shailendra Singh, Neelam Goel and Trilok Chand Aseri , " A Review of Soft Computing Techniques for Gene Prediction", Department of Computer science and Engineering, February 2013 .
[5].   Mingvi Wang, Ping Wu and Shunren Xis," Improving Performance Of Gene selection By Unsupervised Learning",IEEE international conference on Neural networks and signal Processing,2003.
[6].   Yuting Chen, Prof. Er Meng,"Biomedical Diagnosis and Prediction using Parsimonious Fuzzy Neural Networks", Joo School of EEE, Nanyang Technological University Singapore 639798, Singapore, 2012.
[7]   Martin Maiers, Vanja Pauni´c, Michael Steinbach, Vipin Kumar," Prediction of HLA Genes from SNP Data and HLA Haplotype Frequencies", Bioinformatics Research National Marrow Donor program Minneapolis, MN,2012.
[8].   http://www.nature.com/nrd/journal/v4/n1/fig_tab/nrd1611_F1.html
[9].   http://www.redorbit.com/news/science/1112650026/epigenetics-changes-genes-in-rheumatoid-arthritis/