ISSN 0975-2366
https://doi.org/10.31838/ijpr/2018.10.04.019

Review Article

AN EFFECTUAL FILTER BASED GENE SELECTION WITH DNCM-IPSO ALGORITHM FOR DIAGNOSIS OF PERIPHERAL BLOOD CELLS(PBCs) IN RHEUMATOID ARTHRITIS (RAs)

B. CHITHRA[1], DR.R. NEDUNCHEZHIAN[2]
[1]HOD, Department of Computer Technology, Shri Nehru Maha Vidhyala College of Arts and Science, Malumachampatti, Coimbatore. E-mail:chithra220@gmail.com
[2]Professor, Department of CSE & IT, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu. E-mail:nedunchezhian@cit.edu.in, rajuchezhian@gmail.com
Received: 9.07.18, Revised: 9.08.18, Accepted: 9.09.18

ABSTRACT
Rheumatoid Arthritis (RA) is an illness of chronic inflammatory arthritis. Presently, diagnosing RA may involve number of weeks, and the factor applied for the prediction of a poor prognosis is not trustable at all time. Gene expression in RA may contain a distinct signature. Gene expression analysis has been used for synovial tissue for defining molecularly unique forms of RA; but, the expression analysis of tissue obtained from a synovial joint is intrusive and clinically not feasible. The research carried out recently have shown that distinct gene expression variations can be found in Peripheral Blood Mononuclear Cells (PBMCs) taken from different patients affected with cancer, multiple sclerosis and lupus. In order to find RA disease-associated genes, a gene selection and classification was performed. At first, with the result of reducing the time complexity, this dataset related to the illness is procured and then the data is preprocessed. Subsequently, to reduce the number of genes, the gene data is selected from the preprocessed data with the help of filter dependent gene selection methods, which include: T-test, chi-squared test, relief-F and Minimum Redundancy Maximum Relevance (mRMR). Thirdly, Enhance Entropy with Gaussian Kernel based Support Vector Machine (EEGK-SVM) approach is proposed for disease prediction, in turn, maximizes the prediction accuracy. At last, for the RA disease classification, a Dynamic Neutrosophic Cognitive Map with Improved Particle Swarm Optimization (DNCM-IPSO) algorithm is introduced, which is quite suitable for the medical routine and it is presented for aiding the gene expression in the early and accurate diagnosis of RA patients. Consequently, RA disease is not allowed from getting into advanced stages and the hardship with developing insistent and also erosive arthritis for RA patients will also be decreased. The results prove that the DNCM-IPSO methodology performs better performance in terms of precision, accuracy, recall and F-measure etc when compared with the other algorithms.

Key words: Rheumatoid Arthritis (RAs), gene expression profiling, Gene selection, filter based gene selection, Enhance Entropy with Gaussian Kernel, Support Vector Machine (EEGK-SVM), Dynamic Neutrosophic Cognitive Map (DNCM), and Improved Particle Swarm Optimization (IPSO).

INTRODUCTION
Rheumatoid Arthritis (RA) is a disease that is chronic, systemic autoimmune inflammatory with primary joint manifestations [1]. Joint involvement begins with synovial hyperplasia that gradually advances to synovitis and tissue formation, a hugely cellular inflammatory tissue [2]. Tissue leads to the erosion in the cartilage and bone, resulting in articular damage and ankylosis [3]. Articular involvement in the hand and wrist becomes the visible site in nearly 70% of the patients affected with RA and often shows the overall disease condition of the patient [4]. Anti-tumor necrosis factor-alpha (anti-TNFa) referred as infliximab, is basically a chimeric monoclonal antibody, which prevents the proinflammatory cytokine tumor necrosis factor-alpha (TNFa) [5]. But, just a less percentage of patients attain a drastic response (ACR70) and a considerable ratio of patients are not responsive at all to TNF obstruction [6] that is constant with the heterogeneous characteristic of the RA phenotype. Genome-wide gene expression profiling has been utilized for the better classification of several cancers [7][8][9] and to get an understanding about the molecular pathways that involved in various disease processes. Gene expression profiling may permit for an early diagnosis, help in the identification of factors, which provide the prediction of poor prognosis, and assist in focusing on prior, combative, and costly therapy to those patients, who would be the most benefitted. Expression analysis of tissues obtained at the disease site within a synovial joint is intrusive and not feasible as a day to day routine. Nonetheless, the studies carried out recently have confirmed on distinct gene expression variations in peripheral blood mononuclear cells (PBMCs) from patients affected with cancer, multiple sclerosis, and lupus [10].In the last few years, several researchers have noticed that

microarray gene expression data has a significant role to play in the disease diagnosis aiding the doctors in choosing beforehand the necessary treatment plan for the affected patients. One primary challenge faced in biomedical studies in the last few years is the classification of the samples into groups either disease or not-disease. Genetic information obtained in the form of microarray data can be utilized for the identification and classification of a sample belonging to a new patient into disease or no disease classes. But, taking the number and complexity involved with the gene expression data into account, it is quite difficult to get it analyzed manually. Therefore, computational techniques are highly necessary. This is a problem of classification in which task is about learning a classification model depending on a set of labeled training samples obtained from the two populations, and thereafter applying the learned model for the prediction of the label of test samples. Every sample is denoted by numeric values acquired from some kind of gene expression measurements. Understanding the gene expression data is not an easy task due to the high dimensional nature of gene expression data, and also its low sample size. Gene selection is a critical sub problem in such research works. Commonly, the gene microarray data has very less amount of samples, but very huge numbers of genes. Also, it is found that a high number of genes do not have relevance or are repetitive and hence are not helpful in the disease classification. Just a very less number of genes may show relevance [11]. In order to design a rapid system used for the classification of RA disease accurately, feature or gene selection methods are necessary obviously. The over-fitting problems frequently occur if the data is noisy with repetitive features and it comprises of less number of features. Therefore, feature selection can considerably decrease the computational difficulty associated with the task of classification [12]. In order to resolve the above mentioned challenges, in this technical work, four kinds of filtering approaches are introduced. This research work tries to get over the above disadvantages by evolving a prediction and decision support model used for the diagnosis of RA in the early stages employing Improved Particle Swarm Optimization and the soft computational method of Dynamic Neutrosophic Cognitive Map referred to as DNCM-IPSO. These research works are focused on the attempts made to resolve issues in conventional PSOs, and introduces an enhanced PSO model that addresses dynamic behaviors and nonlinear associations in systems. As a result, RA disease can be avoided from getting into progressive stages and the risk of being affected with RA along with PBMC and RAs for these patients would be reduced by making use of DNCM-IPSO. The proposed work provides the prediction of the disease class by making use of EEGK-SVM approach to enhance the prediction accuracy. Then the measurement of the results are performed employing the classification metrics such as precision, recall, F-measure and accuracy and then matched up with the other clustering techniques including Dynamic Neutrosophic Cognitive Map with Bat Algorithm (DNCM-BA), FCM–Particle Swarm Optimization (FCM-PSO), Fuzzy C Means (FCMs), Dynamic Firefly Algorithm Fuzzy C Means (DFAFCM) and Dynamic Fuzzy C Mean (DFCM) clustering algorithms. These algorithms were realized with the help of the MATLAB simulation environment.

Literature Review

The earlier methods of the gene expression profiling in RA has been studied in this section. Chithra & Nedunchezhian [13] explained about the RA consequences on the universal gene expression profile in PBCs involving responder RA. The differentiating gene expression signatures of African American RA patients have been described. The research technique introduces a new Dynamic Fuzzy Cognitive Map (DFCM) with Firefly Algorithm (FA) referred to as DFAFCM algorithm. Dynamic weights are incorporated in FCM model and the trend-effects to generate the model offer more meaning. As a result, a Fuzzy Sigmoid Kernel (FSK) function is designed for weight learning. The results are measured by using the classification metrics including recall, precision, F-measure and accuracy while carrying out the comparison with the earlier clustering methodologies, for example, FCM and DFCM algorithm. Meugnier et al., [14] mentioned the impacts of anti-TNF-α treatment on the global gene expression profile in Peripheral Blood Mononuclear Cells (PBMCs) of responder RA patients. Variations in gene expression were decided employing oligonucleotide microarrays (25,341 genes) in PBMCs acquired before and after 12 weeks of treatment with either etanercept obtained from responder RA patients. Two hundred fifty-one genes showed considerable differences (false discovery rate $< 0.1\%$) in expression level (178 up regulations with mean fold change = 1.5 and 73 down regulations with mean fold change = -1.50) once the 12 week of treatment was over. Universally, inflammation, immune response, apoptosis, protein synthesis, and mitochondrial oxido-reduction were the most impacted pathways as a response to anti-TNF-α treatment. The acquired gene expression signature in PBMCs yields new knowledge in order to better interpret the mechanics of the action pertaining to anti-TNF-α treatment in RA patients. Burska et al., [15] evaluated the presently available information with regard to RA diagnostic, prognostic and prediction of the response obtained to therapy with the objective to focus on the predominance of data, the comparison of which mostly doesn't yield any results owing to the combined usage of material source, experimental techniques and analysis tools, emphasizing the necessity for harmonization in case

gene expression signatures go on to develop as a resourceful clinical tool in customized medicine for RA patients. Woetzel et al., [16] analyzed three multicenter, genome-wide transcriptomic datasets obtained from a total of 79 persons, which were utilized for getting the rule-based classifiers to differentiate RA, Osteoarthritis (OA), and healthy controls. In every case, the rule sets were acquired individually from one among the three centers and used for the other centers for validation purposes. This new approach achieved a superior performance (approximately 90% for specificity, sensitivity, and accuracy) for the RA discrimination. Huo et al., [17] introduced a system for the automated quantification of radiographic finger joint space width of patients affected with early RAs. Nearly 99% of joint locations are identified with an error lesser than 3 mm with regard to the manually specified gold standard. The joint margins are identified by integrating the intensity values and spatially restrained intensity derivatives that are refined by means of an active contour model. Close to 96% of the joints are delineated with success. Aletaha et al., [18] studied about novel classification criteria used for RA, denoting the conclusion of a global coordinated effort aided by both a data-oriented and a consensus-based strategy. Novel paradigm is proposed for the entity "RAs"—specifically, not the criteria for "early" RA. In case there was a treatment, which was both indefinitely efficient and safe and could be rendered at no expense and uncomfortableness, then there would be no need for a subset such as this to be found, since all the patients affected with inflammatory arthritis would get treated. Zheng & Rao [19] introduced the combination of Genome-Wide Association Studies (GWAS) and public knowledge databases to look out for potential pleiotropic genes related to RA and eight other relevant diseases. In this, a GWAS-based network analysis is exploited for identifying the risk genes primarily related to RA. These RA risk genes are then re-acquired in the form of probable pleiotropic genes in case they have been confirmed to be vulnerable genes for at the least one of eight other diseases present in the PubMed databases. Solus et al., [20] analyzed the hypothesis that the mediators of inflammation that known to be in elevated state in Systemic Lupus Erythematosus (SLE) and RA are related to genetic polymorphism earlier detected in the research works involving the inflammatory disease. Genotypes were decided for 345 SNP markers present in 75 genes. Relationship between serum analyses and single alleles was then tested with the help of linear regression. Few genetic relationships are more obvious in healthy controls compared to SLE or RA, indicating dysregulation of the primary mediators of chronic inflammation in disease. Susceptibility genes may have an effect on the inflammatory response with differential effect over the disease etiology. Shiezadeh et al., [21] suggested a predictive model, which provides the diagnosis of RAs. In this research work, a novel classification algorithm called as CS-Boost is proposed; this uses the Cuckoo search algorithm (CSA) for the optimization of the performance obtained of Adaboost algorithm. The dataset involving RAs was obtained from 2,564 patients advised to rheumatology clinic. For every patient, a record comprises of different clinical and demographic features saved. The results obtained from the experiments confirm that the CS-Boost algorithm improves the accuracy of Adaboost in the prediction of RA.

Proposed Methodology

A new DNCM-IPSO and EEGK-SVM prediction model is proposed for the RA disease classification. The entire process of this approach is shown in figure 1. Based on the newly introduced DNCM-IPSO algorithm, dynamic weights are used in NCM model and the trend-effects to make the model to be more meaningful. The learning objective of DNCM is to change the adjacency matrices depending on the meta-heuristic knowledge of the experts that result in the DNCM to get converged into a steady state or desirable area for the target issue. Figure 1 briefly demonstrates an eight-step procedure. (1) Selection of the datasets, (2) identification of RA with gene expression profiles, (3) implementation of the preprocessing model (4) Gene selection process, (5) prediction of the disease employing EEGKL-SVM, (6) implementation of the DNCM model, (7) learning DNCM algorithm with the help of IPSO, (8) evaluation of the learned DNCM-IPSO and (9) analysis of the results. The steps 1, 2 and 9 need the intervention of humans, though the steps 3–7 don't. Also, steps 3–7 are DNCM-IPSO dependent. The results obtained from the experiments corroborate the practicability of the dynamic NCM model.

Selecting the experts' team

Aspects responsible for radiographic difficulties of RA in African-Americans are interpreted badly. It is necessary to examine the genes whose expression in Peripheral Blood Mononuclear Cells (PBMCs) is associated with the radiographic difficulties of RA. 20 control samples (from persons not affected with RA) were compared with 10 early severe, 10 early mild, 10 late mild, and 10 late severe RA samples.

Preprocessing Methods

Reading the dataset samples has been found to be a hugely difficult task. With the objective of solving this problem, few data transformation methodologies are required for reducing the time complexity. Normalization [22] is used in data mining system in the form of a data preprocessing tool. One attribute of a dataset is then normalized through the scaling of its values with the intent that they lie within a small-defined range, for example, 0.0 to 1.0. It is predominantly useful for learning algorithms. The technique of data normalization [22] consists of z-

score normalization, min-max normalization, and normalization by decimal scaling. However, min-max normalization considers min and max values of attributes for normalization. Considering the max values every time does not yield accurate normalization results for solving this problem and instead utilizing maximum value mean value is calculated in this research. The preprocessing steps followed are from the reference [13].

Gene selection process

Primarily, Gene selection methods are introduced for two causes:

Reduce the search space by removing the unnecessary variables

To maximize the predictive potential of the classifiers in supervised learning.

Gene selection refers the procedure of preprocessing the input dataset with the objective of assessing the attributes available with the intent that only the genes related data are kept and irrelevant gene data are removed. Gene selection is useful if the dataset dimensionality is high (where the amount of attributes is many). The popular gene selection techniques such as (T-test, chi-squared test, relief-F and mRMR) are studied in this section.

T-test

This is one among the most widely employed filter techniques for feature selection. With this technique, the statistical importance of the difference of a certain feature between the two classes gets measured [23]. A t-Test is basically a statistic, which checks if the two mean of genes are fundamental diverse from one another. A big t-value indicates diverse groups and a small t-value indicates identical groups. For every t-value, there is a respective p-value. The p-value refers to the probability that the pattern of data present in the sample could be generated with the help of random information. It is observed from experiments that p-value must be <=0.05, which means less than or equal to 5% possibility, then no actual difference exists. Genes having the highest t-statistics are then chosen. T-statistic for every gene is then calculated employing equation 1.

$$t-test = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \quad (1)$$

Where μ_i^+ and σ_i^+denote the mean and standard deviation of values of the i^{th} gene that belong to positive class while μ_i^- and σ_i^-denote the mean and standard deviation of values of the i^{th} gene that belong to negative class. n^+ and n^-denotes the positive sample size and negative sample size correspondingly.

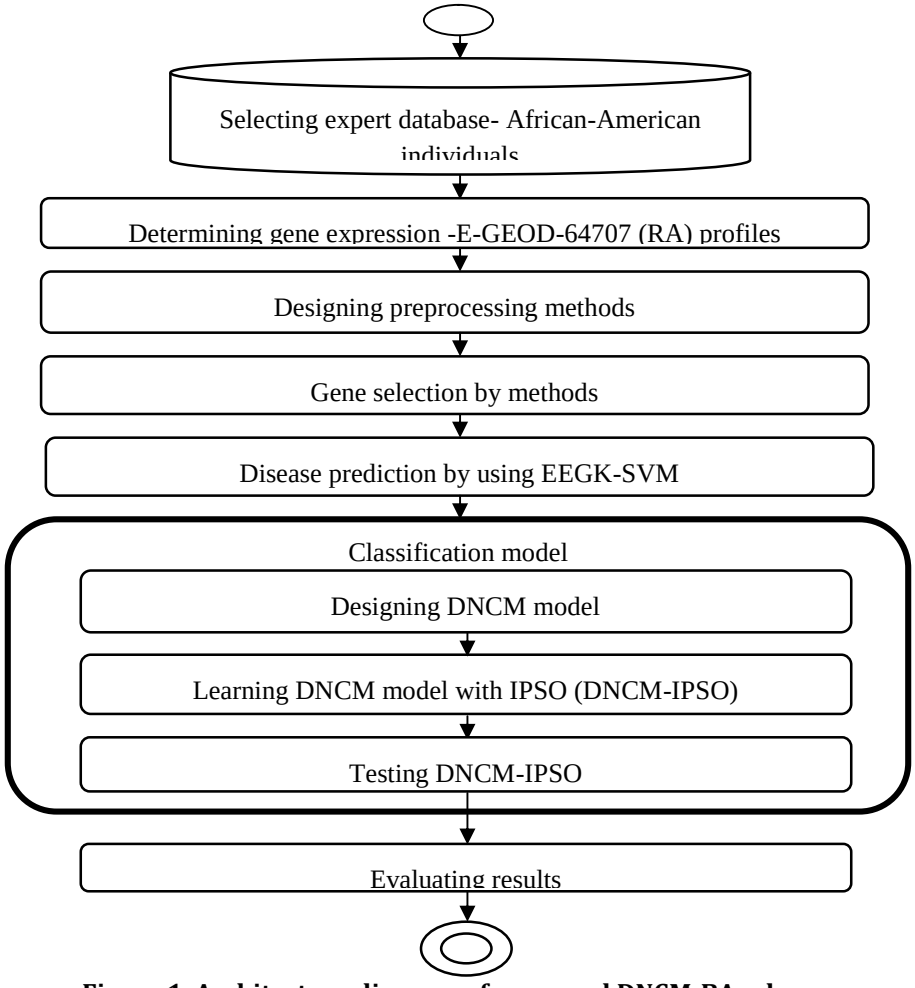

Figure 1. Architecture diagram of proposed DNCM-BA scheme

Chi-squared test

This is also a well-known filter technique, which can be utilized for gene selection. The value of χ^2-Statistic is computed for every individual gene corresponding to the classes [24]. Every numeric gene has to get discretized prior to the computation of χ^2-Statistic. For every gene X_i, χ2-Statistic is expressed as

$$\chi^2 = \sum_{x \in X_i} \sum_{c \in C} \frac{\left(n_{(x \in X_i \& c \in C)} - e_{(x \in X_i \& c \in C)}\right)^2}{e_{(x \in X_i \& c \in C)}} \quad (2)$$

Where $n_{(x \in X_i \& c \in C)}$ refers to the number of samples or patients of class c in which the value of X_i is x. The anticipated frequency $e_{(x \in X_i \& c \in C)}$ is expressed as

$$e_{(x \in X_i \& c \in C)} = \frac{n_{x \in X_i} * n_{c \in C}}{n} \quad (3)$$

Where $n_{x \in X_i}$ represents the number of samples in which X_i has the value x and $n_{c \in C}$ indicates the number of samples of class c. n refers to the total number of samples. The genes are chosen once the ranking of every gene is done according to χ^2-Statistic values.

a. Relief-F

This is an easy and effective technique exploited as a feature (i.e. gene) subset selection technique, to evaluate the quality of the genes, which exhibit extremely huge dependencies between the genes [25]. Relief-F is popular as a filtering technique, which can tackle with noisy and non-uniform datasets. The important concept behind the Relief-F is to evaluate the quality of genes depending on their values to distinguish amongst instances, which are close with one another. With a randomly selected instance Ins_m obtained from class L, the Relief-F calculates the closest hits H that are the K nearest neighbors selected from the same class and also calculates the closest misses M that, in turn, is the K closest neighbors from every one of the diverse classes. It provides the measure of the estimated quality W_i for every gene i based on factors including Ins_m, H, M. In case the instance Ins_m and the others in H have different values on gene i, then W_i is decreased. On the contrary, in case the instance Ins_m and those in M have different values on the gene i, then W_i is raised. The whole process is repeated n times where the value of n is fixed by the users. In order to update W_i, the Equation 4 is computed as below:

$$W_i = W_i - \frac{\sum_{k=1}^{K} D_{HK}}{nK} + \sum_{c=1}^{C} P_c \sum_{k=1}^{K} \frac{D_{MCk}}{n_c K} \quad (4)$$

Where, n_c refers to the no. of instances having class c, n refers to the no of instances with class of Ins_m, P_c stands for the Probability of class c, D_{HK} refers to distance between value of i^{th} gene of k^{th} Hit sample and Ins_m, D_{MCk} stands for the distance between value of i^{th} gene of k^{th} Miss sample of class c and Ins_m.

b. Minimum Redundancy Maximum Relevance (mRMR)

This filter technique chooses the genes having the greatest relevance and minimal redundancy with the target class [26]. In the case of mRMR, the Maximum Relevance and Minimum Redundancy of genes are calculated on the basis of mutual information [26]. Provided the i^{th} gene g_i and the class label c, the mutual information of g_i and c is computed in terms of their probabilities $p(g_i)$, $p(c)$, and $p(g_i, c)$ as given below:

$$I(g_i, c) = \sum \sum p(g_i, c) * ln \frac{p(g_i, c)}{p(g_i) p(c)} \quad (5)$$

The Maximum Relevance technique chooses the greatest top m genes having the most relevance associated with the class labels from the descent ordered set of $I(g_i, c)$.

$$\max D(S, c), D = \left(\frac{1}{|S|}\right) \sum_{g_i \in S} I(g_i, c) \quad (6)$$

Equation (6) depicts the mutual information existing between gene and the class. Nonetheless, researchers know well that "the m best features are not the best m features", as the correlations existing among those top genes may also be huge [26]. Hence, Minimum-Redundancy criterion is presented by [26] with the aim of removing the features with redundancy. The below equation defines the Minimum Redundancy criterion:

$$minR(S), R = \left(\frac{1}{S^2}\right) \sum_{(g_i, g_j) \in S} I(g_i, g_j) \quad (7)$$

Equation (7) indicates that the mutual information between every pair of genes is also taken into consideration. The mRMR filter integrates both the criteria of equations (6) and (7) to compute one single optimization criterion expressed below:

$$max\ \Omega(D, R), \Omega = D - R. \quad (8)$$

An algorithm, which functions in an ordered incremental fashion when resolving the abovementioned optimization criterion is described as below: Suppose, Consider G that denotes a set of genes and also S_{m-1}, which is the gene set containing $m-1$ genes, and so the task is about selecting the m^{th} gene from the set $\{G-S_{m-1}\}$. This is carried out by greedy selection of the feature, which increases Ω. The incremental algorithm consists of a gene g_i to the earlier incrementally generated set S_{m-1} on the basis of the function given below:

$$\max_{g_j \in G - S_{m-1}} \left[I(g_j; c) - \frac{1}{m-1} \sum_{g_i \in S_{m-1}} I(g_j, g_i) \right] \quad (9)$$

According to the above mentioned processes, the prediction of the disease is carried out by making use of EEGK-SVM for boosting the classification accuracy.

Disease prediction using eegk-svm

On a broader perspective, RA disease classification is defined to be the procedure of extraction of distinguished classes (i.e. normal and disease) from the data. It is not unusual that the same type of disease on ground may possess diverse features in PBMCs data. In addition, various types of disease may have identical gene that renders it very difficult to get accurate results either by making use of the conventional unsupervised classification or supervised classification. In this research work, a procedure involving machine learning classifier is used. Support Vector Machine (SVM) involves machine learning used for classification of the types of disease. In this thesis, Enhanced Entropy Based Gaussian Kernel SVM (EEGK-SVM) classifier is introduced for predicting the samples, which are disease based. In this EEMK-SVM classifier, the input feature vectors are first mapped onto a high-dimensional space by means of a nonlinear mapping and after this; a hyperplane is generated and is moved till a suitable separation is attained by the hyperplane with the aim of leaving the maximum possible margin from both the classes. Therefore the objective is to build such a hyperplane out of the training RA disease samples. Provided that $T = \{(f_1, o_1), \ldots, (f_n, o_n)\}$ refers to the training set, $f_i \in R^n$ stand for the input vectors and $o \in \{-1, 1\}$ indicate the RA disease classification labels, the required hyperplane is computed as below,

$$\omega^T f + b = 0 \quad (10)$$

SVMs needs the solution of the optimization problem given below

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} \xi_i \quad (11)$$

$$\text{S.T. } o_i(\omega^T f_i + b) \geq 1 - \xi_i \quad (12)$$

While the actual problem is stated in a finite feature space, it is a frequent happening that in such a kind of space, the RA data, which are supposed to be classified are linearly not differentiable. Therefore it was suggested that the actual finite dimensional space be mapped onto a much greater dimensional space, in which the classes can be isolated by a hyperplane satisfactorily. Kernel functions $\kappa(f, z)$ are computed suitable for the problem.

$$\kappa(f, z) = \sum_{r} \varphi_r(f) \varphi_r(z) \quad (13)$$

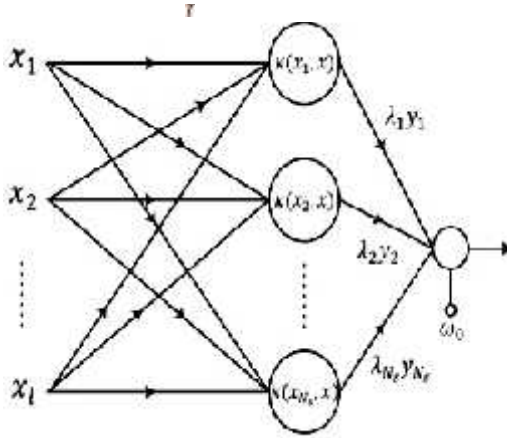

FIGURE 2. Architecture of EEGK-SVM scheme

Figure 2 illustrates the associated architecture ([27-28]).

Gaussian Kernel

The Gaussian kernel is expressed by the equation

$$k(f,z) = \exp\left(-\frac{\|f-z\|^2}{2\sigma^2}\right) \quad (14)$$

It is also called as RBF since it is only based on the distance of f and z present in the input feature vector space. The parameter 'σ' functions similar to degree 'd' in the polynomial kernel managing the kernel's flexibility, where the small values of 'σ' is with respect to the bigger values of 'd'. In case 'σ' is selected too small, the risk of over fitting is increased, whereas a big value of 'σ' decreases the kernel to a constant function that is not suitable for the learning of any essential classifiers. When the kernel function κ(f, z) is adapted for the optimization task, (11) now changes to (15)

$$\max_{\lambda}\left(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j o_i o_j f_i^T f_j\right)$$

$$\text{S.T. } 0 \le \lambda_i \le C, i = 1,\ldots N \quad (16)$$

$$\sum_i \lambda_i o_i = 0 \quad (17)$$

Where, $\lambda_i \geq 0$ refers to the Lagrange multipliers. Once the optimization problem is solved, the resultant linear classifier is computed as below:

$$g(f) = sgn\left(\sum_{i=1}^{n}\lambda_i o_i H\left(k(f_i,z)\right) + \omega_i\right) \quad (18)$$

As observed in equation (44), the outcomes of different kernels are integrated into a new function by making use of the entropy function. Out of these n kernel values, the expected information content H referred as entropy is computed by weighing the kernel values by their corresponding probabilities.

$$H\left(k(f_i,z)\right) = -\sum_{i=1}^{n} P(k(f_i,z))\log_2 P(k(f_i,z)) \quad (19)$$

On the basis of the abovementioned process, the disease prediction and the phase of disease is classified employing the DNCM-IPSO algorithm.

Designing Dncm Model

A Neutrosophic Cognitive Map (NCM) is called as a neutrosophic directed graph (which indicates that a neutrosophic directed graph is actually a directed graph where in as a minimum of one edge is an in determinacy represented by dotted lines) having perceptions like policies, events and so on., to be the nodes and causalities or in determinates to be the edges. In this research work, it represents the causal correlation amongst the RA notions.

Neutrosophic logic

The concept of only fuzzy cognitive maps are handled that mostly handles the relation / non-relation amid two nodes or ideas , but it is quite unsuccessful in managing the association existing among two conceptual nodes whereas the association is an unsure one. Neutrosophic logic is the only popular tool, which deals with the indeterminacy concepts. It is basically a logic where in each proposition is guesstimated so as to comprise the proportion of truth present in a subset T, the proportion of indeterminacy in a subset I, and the proportion of falsity in a subset F, where T, I, F signify the standard or non-standard original subsets. Suppose $C_1, C_2, \ldots\ldots, C_n$ to denote n nodes and the nodes indicate the descriptive RA notions, which could be the features or else behaviors of the system. In addition, each node is actually a neutrosophic vector obtained from neutrosophic vector space V. As a result, a node C_i is represented by $(x_1, x_2, \ldots\ldots, x_k)$ in which x_k's are either 0or 1 or I (I refers to the indeterminate) and $x_k = 1$ indicates that the node C_k is in the greater phase of RA and $x_k = 0$ states that the node is in the lower phase of RA and $x_k = I$ represents the state of the nodes in an intermediate stage of RA.

Dynamic NCM Model

With the objective of broadening the capability of NCMs, a Dynamic NCM (DNCM) model is presented, which is capable of reflecting the dynamic conducts and developing nonlinear associations in systems. For the adjacency matrix $N(E)$ value to concept RAC_i, its relative position finds its specific weight to other notions. Consequently, W_{ij} is expressed based on these grounds:

$$W_{ij} = \begin{cases} 0 \in domain(low) \\ 1 \in domain(large) \\ -1 \in domain(no) \\ I \in domain\ (intermediate\ state) \end{cases} \quad (20)$$

Where W_{ij}vigorously considers three different values. Weight learning of NCM is equivalent to the optimization problem of the connection matrix, which is solved by the equation (20). NCM learning is focused on the learning of the adjacency matrix (X_i) and on the available historical raw RA data. The learning methodologies for NCMs are focused on the learning of the adjacency matrix depending on expert heuristic information or on the available historical RA or on both of them. Evolutionary DNCM learning techniques compute the adjacency matrices from historical data, which is a best fit for the input state vectors sequence. The learning goal of DNCM evolutionary learning is to generate ideal adjacency matrices for individual NCMs modeling particular systems. The learning goal is to modify/bring adjacency matrices from the initial heuristic expert's information and past data with a multi-stage learning process. In this research, the IPSO algorithm is brought into use.

Improved Particle Swarm Optimization (IPSO)

In the case of the standard PSO algorithm [29], the convergence speed of particles is rapid, but the adjustments made of cognition component and social component tend to make the particles to look around for the two best points, which are P_{gb} and P_{lb}. Based on the velocity and the position updating formula, when the best individual in the swarm gets trapped into a local optimum, the information sharing strategy in the standard PSO will have other particles attracted in order to near this local optimum slowly. Finally, the entire swarm will get converged to the local optimum. The chief reason behind the standard PSO algorithm to get easily slipped into local optima is that it contains no techniques to enforce the particles to move out of the local optima. This is PSO's critical drawback. This critical setback forms the reason on why PSO could not always attain the optimal solution. With this suggestion in this research work, a novel algorithm is introduced in the following subsection that pushes the PSO to prevent the slipping into local optima. An enhanced Version of PSO by taking the levy operator into consideration is added to the actual PSO algorithm. The important objective of IPSO is to prevent the slipping into the local optima. With the advice on the primary objective, the IPSO adapts for getting the optimal adjacency matrix. This adjacency matrix tends to make the particles to move on the opposite direction of the worst adjacency matrix positions and the worst entire swarm positions. Therefore, it expands the global searching space and limits the probability of the particles to slip into a local optimum.IPSO is required to generate the pessimistic flying direction for every particle to choose the optimal set of adjacency matrix out of the RA disease data. The process of generating the pessimistic flying direction is similar to that of generating the optimistic flying direction. The standard PSO records the best selection an adjacency matrix (particle) achieved by it, and the best ones experienced by the entire swarm. Depending on the best position of every adjacency matrix (particle) and the best position of the entire swarm, the standard PSO generates the optimistic direction employing the Formula (21) and (22).

$$V_{id}^{'} = \omega V_{id} + \eta_1 rand(P_{ib} - X_{id}) + \eta_2 rand(P_{gb} - X_{id}) \quad (21)$$

$$X_{id}^{'} = X_{id} + V_{id}^{'} \quad (22)$$

Where ω is known as the inertia weight. It is actually a proportional factor related with former velocity . η_1 and $\eta_2$stand for constant accelerating factors, generally $\eta_1 = \eta_2 = 2$. The random function rand () is used for the generation of random numbers. X_{id} denotes the position of adjacency matrix id. V_{id} refers to the velocity of adjacency matrix id. P_{ib} and P_{gb} stands for the best position of the adjacency matrix id found and the best position of the entire swarm observed correspondingly until this moment.To generate the pessimistic direction, the process of generating the optimistic direction employing in the standard PSO can be modified. It means that, the worst position achieved by an individual adjacency matrix (particle), and the worst ones experienced by the entire swarm can also be recorded. Depending on the worst position of every adjacency matrix (particle) and the worst position of the entire swarm, IPSO generates the pessimistic direction with the Formula (23) and (24) that specifies the adjacency matrix velocity and position updating as below:

$$V_{id}^{'} = \omega V_{id} + \eta_1 levy(X_{ib} - P_{idw}) + \eta_2 levy(X_{id} - P_{gdw}) \quad (23)$$

$$X_{id}^{'} = X_{id} + V_{id}^{'} \quad (24)$$

Where P_{idw} denotes the worst position found by the adjacency matrix id. P_{gdw} represents the worst positions found by the entire swarm. IPSO selects the best one from the two probable flying directions depending on the levy operator to have every adjacency matrix updated. As observed in equation (23), the levy operation is presented to get both the global and local position of the adjacency matrix updated as described below. Levy is presented as a distribution that provides an infinite second moment, dissimilar to the finite second moment in Gaussian distribution, called as Levy's Probability Distribution [30]. It is also a stabilized process having an infinite moment that yields a distinctive tail finally. The distribution is expressed as below.

$$L_{(\alpha,\gamma)}(Y) = \frac{1}{\pi}\int_0^{\infty} e^{-\gamma q^{\alpha}} \cos(qy)\, dq \quad (25)$$

Also, in the experiments, all the algorithms are run 100 times for each function, and the parameters are fixed at $\omega = 0.6$, $\eta_1 = \eta_2 = 2$ The framework of IPSO is described in Algorithm 1. It can be seen from Eqn. (25),

the distribution is symmetric with regard to y=0 and contains two parameters γ and . γrefers to the scaling factor that satisfies $\gamma > 0$ and α must satisfy $0 < \alpha < 2$. Keeping the limits especially for α = 1 into consideration, the integration can be analytically carried out where it equals to Cauchy Probability distributions. For the limit $\alpha \to 2$, the distribution shifts towards Gaussian distribution. This certainly expands the global searching space of the particles, and facilitates them to prevent slipping into a local optimum too before and simultaneously. Therefore, the improved PSO increases the probability of getting the global optimum in the search space.

Algorithm 1: IPSO based wrapper feature selection algorithm

Initialize the velocities $V_i = (V_1, \ldots V_n)$and positions of adjacency matrices $X_i = (X_1, \ldots, X_n)$

Calculate the accuracy of the classifier of everyX_i

For every$X_i$if its accuracy value is smaller compared to the best accuracy value P_{ib}

Update its best position P_{ib}Else if its accuracy value is greater than the worst accuracy value P_{ibw}

update its P_{idw}

For everyX_iIf its accuracy value is lesser than the best of the entire swarm accuracy value P_{gb}

Update the whole swarm best accuracy value P_{gb} employing the accuracy value of this feature

Else if its accuracy value is greater than the worst entire swarm accuracy value P_{g}

$_{dw}$update P_{gdw} employing the fitness value of this particle;

For every feature,

Generate the optimistic X_i selection position t of this feature by formula (21, 22)

Generate the pessimistic potential position t' of this particle by formula (23, 24)

Perform a comparison between t and t' and then choose the better one to update this X_i;

Repeat from Step 2 till the termination criterion is met.

The innovative DNCM with IPSO learning classifier is constructed the reasoning capability further than the scarce available data – the augmentative as well as analytical skill of the human domain specialists for rational decision making in a medical setting on the subject of harshness of RA disease. For the purpose of analysis, several numbers of random patients are considered and the computation of their severities of RA along with gene profiles is carried out with the help of MATLAB tool.

Experimental Resutls

Aspects responsible for radiographic harshness of RA in African-Americans are not interpreted correctly. It is necessary to examine the genes whose expression in Peripheral Blood Mononuclear Cells (PBMCs) is associated with radiographic harshness of RA. 20 control samples (from individuals not affected with RA) were compared with 10 early severe, 10 early mild, 10 late mild, and 10 late severe RA samples. Every sample was acquired from African-American humans. With the intent of explaining the unique expression signatures in African American RA patients [13] with critical erosive disease, a gene expression analysis is considered by using the samples of RNA obtained from PBMCs. The newly introduced DNCM-IPSO dependent unsupervised learning system and existing DNCM-BA, DFAFCM, FCM, FCM-PSO and DFCM classifiers are measured with the help of the classification metrics is implemented in MATLAB. Based on predictive analytics, a table consisting of confusion matrix is defined as table having two rows and two columns that specifies the amount of false positives, true positives, false negatives, and true negatives.

Table 1. Results comparison vs. learning models

Methods	Results (%)							
	FPR	TNR	FNR	Precision (P)	Recall (R)	F-measure	Accuracy	Error
FCM	66.67	33.33	22.22	77.78	77.78	77.78	66.67	33.33
FCM-PSO	50.00	50	17.39	84.44	82.61	83.52	75	25
DFCM	46.66	53.33	11.11	85.11	88.89	86.96	80.00	20
DFAFCM	33.33	66.67	10.42	91.49	89.58	90.53	85.00	15
DNCM-BA	31.21	69.14	9.95	93.24	92.03	91.25	88.63	11.37
DNCM-IPSO	30.18	71.23	9.02	94.02	93.1	92.12	89.12	10.88

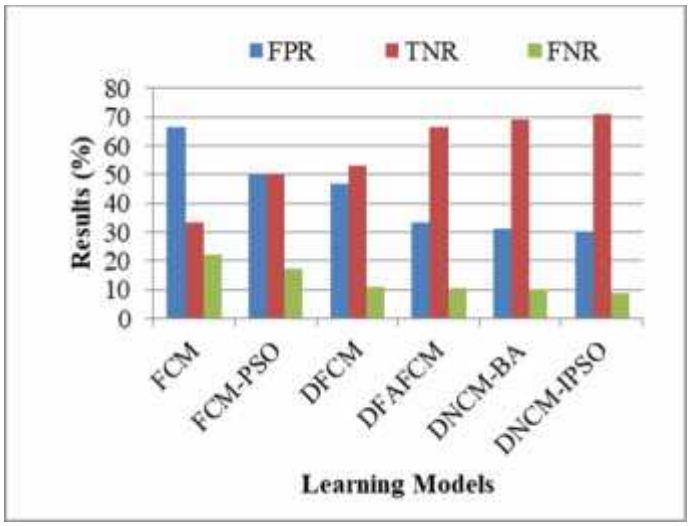

Figure 3. Learning models vs. results (TNR, FNR and FPR)

Figure 3 demonstrates the results of performance analysis of six different classifiers for example FCM, FCM-PSO [11], DFCM DFAFCM [11], DNCM-BA and DNCM-IPSO (proposed) classifier in terms of TNR, FPR and FNR. It shows that the newly introduced DNCM-IPSO classifier renders higher TNR detection percent and lesser FPR, FNR. The FCM classifier renders 66.67%, 33.33% and 22.22% for FPR, TNR and FNR respectively. The FCM-PSO classifier offers 50%, 50% and 17.39% for FPR, TNR and FNR respectively. The DFCM classifier achieves 46.66%, 53.33% and 11.11% for FPR, TNR and FNR respectively. The DFAFCM classifier attains 33.33%, 66.77% and 10.42 % for FPR, TNR and FNR. The DNCM-BA classifier achieves 31.21%, 69.14% and 9.95% for FPR, TNR and FNR respectively. The novel DNCM-IPSO classifier accomplishes 30.18%, 71.23% and 9.02% for FPR, TNR and FNR respectively. The figure 3 shown above proves that the newly introduced DNCM-IPSO offers a lesser FNR rate of 9.02%, which is 0.93 %, 1.4%, 2.09%, 8.37% and 13.2% lesser when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM techniques respectively. It is also shown that the newly introduced DNCM-IPSO renders a lesser FPR rate of 30.18%, which is 1.03 %, 3.15%, 16.48%, 19.82% and 36.49% lesser when compared to DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM techniques respectively. It is proved that the newly introduced DNCM-IPSO renders a higher TNR rate of 71.23%, which is 2.09%, 4.56%, 17.9%, 21.23% and 37.9% higher when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM techniques respectively. It is proven that the newly introduced DNCM-IPSO achieves higher TNR when matched up with other methodologies.

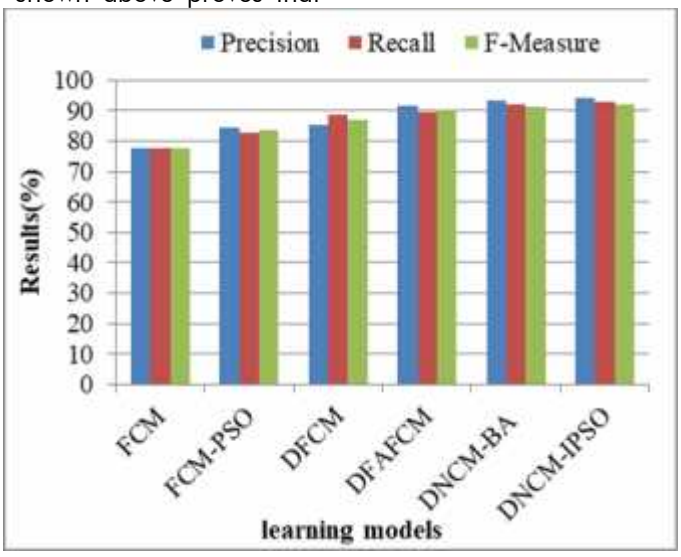

Figure 4. Learning models vs. outcomes (TNR, FNR and FPR)

Figure 4 shows the results of the performance analysis carried out for six different classifiers for example FCM, FCM-PSO, DFCM, DFAFCM, DNCM-BA and DNCM-IPSO (introduced) classifier in terms of recall, precision and f-measure. It depicts that the newly introduced DNCM-IPSO classifier achieves a higher f-measure, precision and recall. The FCM classifier provides 77.78%, 77.78% and 77.78% for

precision, recall and f-measure respectively. The FCM-PSO classifier attains 84.44%, 82.61% and 83.52% for precision, recall and f-measure respectively. The DFCM classifier renders 85.11%, 88.89% and 86.96% for precision, recall and f-measure respectively. The DFAFCM classifier attains 91.49%, 89.58% and 90.53 % for precision, recall and f-measure. The DNCM-BA classifier achieves 91.49%, 89.58% and 90.53 % for precision, recall and f-measure correspondingly. The newly introduced DNCM-IPSO classifier accomplishes 94.02%, 93.1%, 92.12% for precision, recall and f-measure respectively.The figure 4 shown above proves that the newly introduced DNCM-IPSO achieves a higher f-measure rate of 92.12%, which is 0.87%, 1.59%, 5.16%, 8.6%, and 14.34% higher when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM techniques respectively. The newly introduced DNCM-IPSO offers a higher recall rate of 93.1%, which is 1.07%, 3.52%, 4.21%, 10.49% and 15.32 higher when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM methodologies respectively. The newly introduced DNCM-IPSO achieves a higher precision rate of 94.02%, which is 0.78%, 2.53%,8.91%, 9.58%, and 16.24% higher when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM strategies respectively.

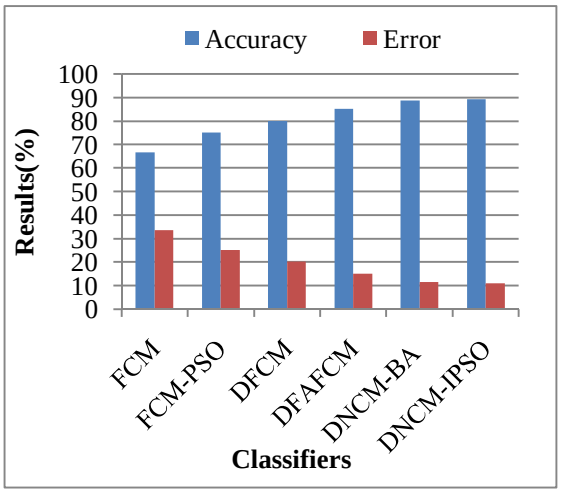

Figure 5. Learning models vs. outcomes (accuracy and error rate)

Figure 5 shows the results of the performance analysis carried out for four different classifiers for example FCM, FCM-PSO, DFCM, DFAFCM, DNCM-BA and DNCM-IPSO (suggested) classifier in terms of accuracy and error rate. It shows that the newly introduced DNCM-IPSO classifier renders superior accuracy, and lesser error value. The FCM classifier offers 66.67% and 33.33% for accuracy and error values respectively. The FCM-PSO classifier achieves 75% and 25% for precision accuracy and error values respectively. The DFCM classifier attains 80% and 20% for accuracy and error values respectively. The DFA-FCM classifier achieved 85% and 15% for accuracy and error values respectively. The newly introduced DNCM-BA classifier produces 88.63% and 11.37% for accuracy and error values respectively. The proposed DNCM-IPSO classifier produces 89.12% and 10.88% for accuracy and error values respectively. It is proved that the newly introduced DNCM-IPSO achieved a higher accuracy rate of 89.12% that is 0.49%, 4.12%, 9.12%, 14.12% and 22.45% higher when compared with DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM methodologies respectively.

Conclusion

Diagnosis of RA in its early stage is not possible. There are no exact test, which can clearly detect RA; therefore a system, which aids in the early prediction and identification of RA beforehand, is highly required. In this research work, a new Dynamic Neutrosophic Cognitive Map with Improved Particle Swarm Optimization (DNCM-IPSO) with EEGK-SVM prediction model is used for getting the gene expression profiles that distinguishes the patients with RA from potential control subjects. In this thesis, at first, the data preprocessing is carried out and then gene is selected through the T-test, chi-squared test, relief-F and Minimum Redundancy Maximum Relevance (mRMR). Thirdly, the prediction of the disease is done with the help of Enhance Entropy with Gaussian Kernel based Support Vector Machine (EEGK-SVM) approach that maximizes the prediction accuracy. Subsequently, the selected gene expressions are sent to learning process. In the DNCM-IPSO learning model, genes are analyzed whose expression in PBMCs is associated with radiographic harshness of RA. Twenty control samples (acquired from persons not affected with RA) were compared with 10 early mild, 10 early severe, 10 late mild, and 10 late severe RA samples. All samples were obtained from African-American individuals. The results of performance analysis carried out for six different classifiers for example FCM, FCM-PSO, DFCM, DFAFCM, DNCM-BA and DNCM-IPSO (introduced) classifier in terms of

accuracy and error rate. It proves that the newly introduced DNCM-IPSO achieves a higher accuracy rate of 89.12% , which is 0.55%, 4.622%, 10.23%, 15.84% and 25.19% more when compared with the earlier DNCM-BA, DFAFCM, DFCM, FCM-PSO and FCM techniques respectively.

References

1. Gravallese, E.M., 2002. Bone destruction in arthritis. Annals of the rheumatic diseases, 61(suppl 2), pp.ii84-ii86.
2. Haavardsholm, E.A., Bøyesen, P., Østergaard, M., Schildvold, A. and Kvien, T.K., 2008. Magnetic resonance imaging findings in 84 patients with early rheumatoid arthritis: bone marrow oedema predicts erosive progression. Annals of the rheumatic diseases, 67(6), pp.794-800.
3. Arthritis Foundation, Disease Center, http://www.arthritis.org/conditions/DiseaseCenter/RA/default.asp. (Accessed: 18 October 2006). 2006.
4. Odeh M. New insights into the pathogenesis and treatment of rheumatoid arthritis. Clin Immunol Immunopathol 1997;83: 103—6.
5. Maini RN, Feldmann M. How does infliximab work in rheumatoid arthritis? Arth Res 2002;4(Suppl 2):S22—8.
6. Weinblatt ME, Kremer JM, Bankhurst AD et al. A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. N Engl J Med 1999; 340: 253–259.
7. Staudt LM. Gene expression profiling of lymphoid malignancies. Annu Rev Med 2002; 53: 303–318.
8. Bhatt P, Khatri N, Kumar M, Baradia D, Misra A. Microbeads mediated oral plasmid DNA delivery using polymethacrylate vectors: an effectual groundwork for colorectal cancer. Drug delivery, 22(6), 2015, 849-61.
9. Patil S, Bhatt P, Lalani R, Amrutiya J, Vhora I, Kolte A, et al. Low molecular weight chitosan-protamine conjugate for siRNA delivery with enhanced stability and transfection efficiency. RSC Advances. 2016;6(112):110951-63.
10. Qing, X. and Putterman, C., 2004. Gene expression profiling in the study of the pathogenesis of systemic lupus erythematosus. Autoimmunity reviews, 3(7-8), pp.505-509.
11. Osareh, A. and Shadgar, B., 2010, Microarray data analysis for cancer classification. 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), pp. 125-132.
12. J. Zhang, and H. Deng. "Gene selection for classification of microarray data based on the Bayes error". BMC Bioinformatic, 8:370, 2007.
13. Chithra, B., & Nedunchezhian, R. (2017). Medical Diagnosis of Peripheral Blood Cells (PBCS) Gene Expression Profiling in Rheumatoid Arthritis (RA) Using DFAFCM Algorithm, Journal of Advanced Research in Dynamical and Control Systems, 15-Special Issue, pp.171-186.
14. Meugnier E, Coury F, Tebib J, Ferraro-Peyret C, Rome S, Bienvenu J, Vidal H, Sibilia J, Fabien N. Gene expression profiling in peripheral blood cells of patients with rheumatoid arthritis in response to anti-TNF- treatments. Physiol Genomics 43,pp. 365–371, 2011.
15. Burska, A. N., Roget, K., Blits, M., Gomez, L. S., Van De Loo, F., Hazelwood, L. D., ... & Ponchel, F. (2014). Gene expression analysis in RA: towards personalized medicine. The pharmacogenomics journal, 14(2), pp.93-106.
16. Woetzel, D., Huber, R., Kupfer, P., Pohlers, D., Pfaff, M., Driesch, D., ... & Kinne, R. W. (2014). Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. Arthritis research & therapy, 16(2), R84.
17. Huo, Y., Vincken, K.L., van der Heijde, D., De Hair, M.J., Lafeber, F.P. and Viergever, M.A., 2016. Automatic quantification of radiographic finger joint space width of patients with early rheumatoid arthritis. IEEE Transactions on Biomedical Engineering, 63(10), pp.2177-2186.
18. Aletaha, D., Neogi, T., Silman, A.J., Funovits, J., Felson, D.T., Bingham III, C.O., Birnbaum, N.S., Burmester, G.R., Bykerk, V.P., Cohen, M.D. and Combe, B., 2010. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis & Rheumatism, 62(9), pp.2569-2581.
19. Zheng, W., & Rao, S. (2015). Knowledge-based analysis of genetic associations of rheumatoid arthritis to inform studies searching for pleiotropic genes: a literature review and network analysis. Arthritis research & therapy, 17(1), pp.1-9.
20. Solus, J. F., Chung, C. P., Oeser, A., Li, C., Rho, Y. H., Bradley, K. M., ... & Stein, C. M. (2015). Genetics of serum concentration of IL-6 and TNFα in systemic lupus erythematosus and rheumatoid arthritis: a candidate gene analysis. Clinical rheumatology, 34(8), 1375-1382.
21. Shiezadeh, Z., Sajedi, H., & Aflakie, E. Diagnosis Of Rheumatoid Arthritis Using An Ensemble Learning Approach.ICAITA, SAI, CDKP, Signal, NCO - 2015 pp. 139–148,
22. Jain, Y.K. and Bhandare, S.K., 2011. Min max normalization based data perturbation method for privacy protection. International Journal of Computer & Communication Technology, 2(8), pp.45-50.
23. Ding, C. and Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), pp.185-205.
24. Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), pp.1289-1305.
25. M. E. Wall, A. Rechtsteiner, and L. M. Rocha. "A Practical Approach to Microarray Data Analysis". Norwell, MA: Kluwel. chapter 5, 91–109, 2003.
26. Peng H et al. "Feature selection based on mutual information: criteria of maxdependency max

relevance, and min-redundancy". IEEE Trans Pattern Anal Mach Intell;27:1226–38, 2005.
27. J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
28. S. Theodoridis and K. Koutroumbas. Pattern Recognition. 3rd Edition, Academic Press, 2006.
29. Bai, Q., 2010. Analysis of particle swarm optimization algorithm. Computer and information science, 3(1), pp.180-184.
30. Applebaum, D., 2009. Lévy processes and stochastic calculus. Cambridge university press.