International Journal of Engineering & Technology, 7 (2.21) (2018) 242-250

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Gene selection and dynamic neutrosophic cognitive map with bat algorithm (DNCM-BA) for diagnose of rheumatoid arthritis (RAs)

B. Chithra[1*], R. Nedunchezhian[2]

[1]HOD, Department of Computer Technology, Shri Nehru Maha Vidhyala College of Arts and Science, Malumachampatti, Coimbatore.
[2]Professor, Department of CSE & IT, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu.
**Corresponding author E-mail:chithra220@gmail.com*

Abstract

Rheumatoid Arthritis (RA) is an autoimmune inflammatory rheumatic disease that has emotional impact on various body parts and tissue, principally the synovial joints. RA is a complex disease similar to many other autoimmune diseases, in which environmental aspects, genetic variants, as well as arbitrary events cooperate to activate pathological pathways. Choosing the appropriate gene for sample classification is very hard in numerous gene expression analyses in RA, in which authors attempt to find out the least probable set of genes, even now which could attain better predictive performance. On the other hand, the accuracy of classification is not up to the mark. As a result, for identifying RA disease, this research presents a gene selection as well as classification technique. Initially, with the aim of decreasing the time complexity, this disease dataset is preprocessed. Next, so as to decrease the amount of gene, the gene data is chosen from the preprocessed data by means of filter based gene selection techniques: Chi-square (CHI), Information Gain (IG), Consistency Based Subset Evaluation (CS) and Correlation Based Gene Selection (CGS). Thirdly, for the purpose of the classification of RA disease, a Dynamic Neutrosophic Cognitive Map with Bat Algorithm (DNCM-BA) is presented, that is well-suited with the medical routine and it is proposed for supporting gene expression beforehand and accurate diagnosis of RA patients. As a result, RA disease is not permitted from moving to progressive phases and the difficulty of emerging insistent as well as erosive arthritis for RA patients will get reduced. Finally, the outcome confirms that the DNCM-BA technique provides better performance while matched up with 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 in regard to precision, accurateness, recall and F-measure.

Keywords: *Gene selection, gene expression profiling, filter based gene selection , Bat Algorithm (BA),Peripheral Blood Cells (PBCs), Consistency Based Subset Evaluation (CS), Correlation Based Gene Selection (CGS) and Dynamic Neutrosophic Cognitive Map (DNCM), Rheumatoid Arthritis (RA).*

1. Introduction

Rheumatoid Arthritis (RA) is a heterogeneous syndrome described initially by chronic inflammation in addition to the damage of bone and gristle in di-arthrodial joints. The reason of RA is mysterious, with genetic as well as environmental aspects be part of the cause to disease susceptibility[1]. Inflammatory rheumatoid synovium is described by primary lining layer hyperplasia as well as angiogenesis united with infiltration by a composite blend of mast cells, T and B lymphocytes, dendritic cells and macrophages[2]. The comparative significance of these cell categories for disease beginning and continuation is indeterminate. T cells as well as Activated monocytes might be identified in peripheral blood and synovial tissues, and these cells are powerful resources of pro-inflammatory cytokines for instance Tumour Necrosis Factor (TNF)-α, which act as a significant role in disease pathogenesis. Present biologic therapies, which neutralize TNF, have exposed a noteworthy medical progress in RA patients. On the other hand, just a lesser ratio of patients attain an intense response as well as a substantial amount of patients don't react little to TNF blockade, [3, 4] steady with the heterogeneous nature of the RA phenotype.

For about 70% of patients, TNF-α treatment brings a noteworthy medical enhancement [5]. On the other hand, the degree of medical enhancement is habitually distant from whole remission as well as numerous RA patients undergo a burst of the disease within the initial 2 years [5]. As a result, for the progress of novel treatments, the finding of novel molecules, which act as an essential role in the pathogenesis of the disease, is important. Additionally prognostic as well as diagnostic biomarkers must be found. So, as it is utilized for categorizing lymphoid malignancies [6] in addition to separate paths in more than a few autoimmune and inflammatory diseases, gene expression profiling provides better assistance. By the way, gene array analysis produced beneficial vision in numerous diseases comprising Systemic Lupus Erythematosus (SLE) [7], RA [8], as well as Multiple Sclerosis (MS) [9].

Practically, each facet of a disease phenotype must be denoted in the form of active gene and consequent transcripts and proteins, which are stated. Deoxyribonucleic Acid (DNA) microarray technology is a dominant method, which assists for the analysis of mRNA levels of the gene in the genome. Use of significant gene expression profiling by utilizing DNA micro arrays of blood as well as tissue samples of RA patients leads an open survey for finding out broadly the ratio of active gene, which are in particular for a medical conditions.

For finding out transcriptional profiles, which differentiate patients with RA from strong control subjects, a gene array approach is utilized in this analysis. So, initially, gene expression in paired Peripheral Blood Cells (PBC) as well as synovial biopsies of RA patients is examined. Finding of particular gene 'signatures', which are diversely stated in patients while matched up with healthy subjects is accomplished by utilizing Gene expression analysis in PBCs of subjects with RA and SLE [8, 9]. As the normal gene selection methods utilize univariate rankings for measuring the gene relevance and it has been used to only two-class problems. However the use of this gene selection ranking criteria to distinct class samples becomes very difficult task. This research presented gene selection algorithm for resolving this issue.

Identifying a gene subset as small as possible is the goal of gene selection. In order to decrease data dimensionality, it eliminates inappropriate as well as redundant genes. Consequently, it increases the mining accuracy and results in comprehensibility and decreases the computational time [10-11]. Make use of the mining steps to the reduce gene subset provides better classification results for real high- dimensional dataset samples. Minimum storage and training times, enabling data visualization which evades overfitting, fast implementation and running of mining algorithms are some of the benefits of Gene selection [12].

A Gene selection process involves four vital steps those are as follows: first one is gene subset generation, second one is subset evaluation, third one is stopping criterion and the final one is result validation. The gene subset generation is a heuristic search process that brings about the choice of a candidate subset for assessment. In the recent work genes are selected by using arbitrary search. By means of an evaluation criterion, the efficiency of the produced subset is assessed. It substitutes the former subset with the finest subset when the recently produced subset is superior to the existing one. Until the stopping criterion is obtained, these two processes are performed repeatedly. After that, through former knowledge or by performing diverse tests, the final finest gene subset is verified. The gene selection process is illustrated in Figure 1.

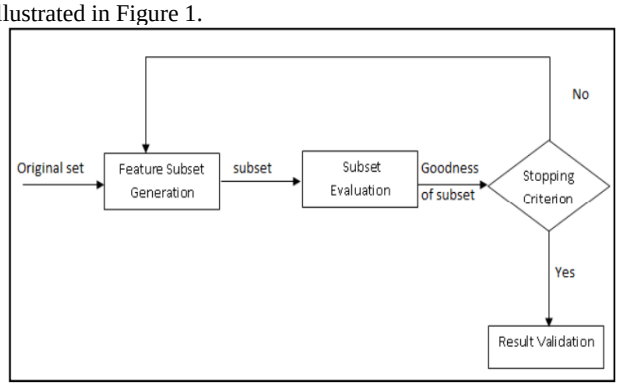

Figure 1: Gene Selection Process

Gene selection algorithms are categorized into three types based on selection approach: Filter [13], Wrapper [14] and Hybrid Method [15]. Based on the intrinsic features of the data, Filter technique chooses the gene subset independent of mining algorithm. It is applied to data having high dimensionality. Its generality and greater computation efficiency are the benefits of Filter method. For identifying the finest gene subset, Wrapper Technique needs a predetermined algorithm. Predictive accurateness of the algorithm is utilized for assessment. This technique provides better outcomes; on the other hand, it computationally costs high for huge dataset. So, it is not generally chosen. With the aim of attaining the benefits of both the approaches, Hybrid Method unites Filter and Wrapper techniques. For evaluating the goodness of recently produced subset, it utilizes a distinct measure and a mining algorithm [16].

The process of filter based gene selection approaches with four metrics is conversed in the part 3. For better knowing the basic mechanisms of RA, the current analysis focused on finding out the important gene in RA. The incorporated examination of expression profiling was carried out to Peripheral Blood Cells (PBC) in RA. Elucidating certain facets of the RA disease and finding out particular gene signatures in paired PBC with RA are the objectives of this research. For gene reduction in order to raise classification accurateness, this research presented gene selection techniques such as CHI, IG, CS and CGS. The extension of fuzzy classifier, which is neutrosophic classifier that is user-friendly decision support tool and would utilize neutrosophic logic that is a superset of fuzzy logic. The DNCM-BA proposed for finding out particular gene signatures. The outcomes are deliberated by means of the classification metrics such as recall, precision, F-measure and accurateness matched up with the previous clustering approaches for instance FCM-PSO algorithm, DFAFCM, FCM and DFCM, which were implemented in MATLAB simulation environment.

2. Literature review

Former techniques of the gene expression profiling in RA has been conversed in this part. Chen et al [17] presented new microRNAs diversely proposed in RA osteoblasts and for finding genes possibly engaged in the dysregulated bone homeostasis in RA. In assessing treatments aiming chemotaxis and neovascularization for controlling joint destruction in RA, the discoveries denote novel candidate genes as the powerful pointers.

Wang et al [18] focused on identifying novel genes related to RA with the intension that more widespread genes will be utilized for observing and diagnosing patients. On pathway analysis as well as protein–protein interaction networks, Bioinformatics was carried out. Pathway analyses exposed 10 meaningfully improved pathways, and a protein–protein interaction network analysis proved that four new PBMC-derived genes were carried out to formerly stated genes by four intermediary genes.

Tchetina [19] proposed the discoveries denote that greater TNF-α-related gene expression in the PB is a precondition of a better response of RA patients to MTX and biological treatment. RA patients having low levels of TNF-α expression, who as well containing the expression of other genes at the degree of strong subjects, must be stratified into a diverse subset needing particular treatment that may include targets except pro-inflammatory cytokine signaling paths.

Yoshida et al [20] assessed gene expression in the micro-dissected synovial lining cells of RA patients, by means of those of Osteoarthritis (OA) patients as the control. The molecular activity of RA was reliable with its medical as well as histological activity.

Suzuki et al [21] examined DNA microarray-based gene expression anomalies in peripheral blood from patients with Systemic Lupus Erythematosus (SLE) and RA. By means of utilizing numerous bioinformatics approaches, these datasets were examined. Additionally, all-inclusive single nucleotide level variant examination found new multiple is forms (SLE: 125, RA: 79) characterized by SLE and RA.

Giannopoulou et al [22] analyzed the aspects, which could control gene expression is of considerable significance for rheumatic diseases with badly understood etiopathogenesis. In this research, converse various advantages of this next-generation sequencing technology for assessing rheumatic disease patients and examine the pathogenesis of rheumatic diseases for instance systemic lupus erythematosus, rheumatoid arthritis, juvenile idiopathic arthritis and Sjögren's syndrome.

Edwards et al [23] proposed periodic assessment of gene expression for diagnosis and observing in RA might give a valuable technique for identifying subclinical disease progression and keep an eye on responses to therapy with Disease Modifying Anti-Rheumatic Agents (DMARDs) or anti-TNF-α therapy. These outcomes recommend that valuation of peripheral blood gene expression might show valuable to observe disease development and react to therapy.

3. Proposed methodology

A novel DNCM with BA called DNCM-BA is proposed in this paper. It is a user-friendly decision support technique for accurate diagnosis of RA patients. As a result, RA disease is prevented from progression and the threat of developing insistent and erosive arthritis for these patients will be reduced. According to the proposed DNCM-BA algorithm, dynamic weights are proposed to NCM model and trend-effects to create the model more sensible. DNCM learning objective is to alter adjacency matrices based on the experts' meta-heuristic knowledge, which lead the DNCM to converge into a steady state or into a suitable area for the target issue. Figure 2 depicts an eight-step procedural outline. (1) Choosing the datasets, (2) identifying RA with gene expression profiles, (3) implementing the preprocessing model (4) gene selection process, (5) implementing the DNCM model, (6) learning DNCM-BA algorithm with FA, (7) assessing learned DNCM-BA, and (8) valuing the outcomes. The steps 1, 2 and 8 require human involvement, however steps 3–7 don't. Moreover, steps 3–7 are DNCM-BA based ones. Experimental results indicate the feasibility of the dynamic NCM model.

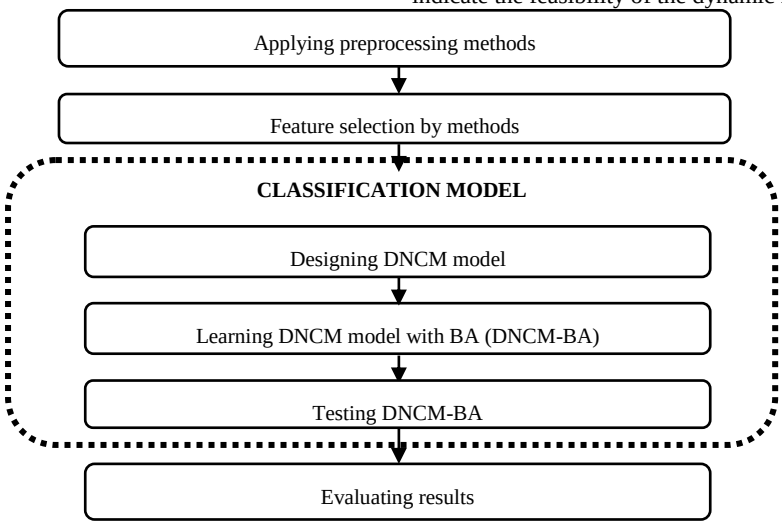

Figure 2: Architecture of the proposed DNCM-BA technique

3.1. Selecting the experts' team

Aspects accountable for radiographic harshness of Rheumatoid Arthritis (RA) in African-Americans are badly understood. So it is essential to select the expert's team to inspect genes whose expression in Peripheral Blood Mononuclear Cells (PBMCs) is related to radiographic harshness of RA. 20 control samples (from individuals not having RA) were matched up with 10 early severe, 10 early mild, 10 late mild, and 10 late severe RA samples.

3.2. Preprocessing methods

The reading and the interpretation of data samples turn out to be an extremely hard task. With the aim of resolving this issue certain data transformation techniques are needed for decreasing time complexity. Normalization [24] is utilized in data mining system as a data preprocessing tool. The Features of a dataset is normalized by scaling its values with the intension that they fall inside a small-specified range, for instance 0.0 to 1.0. The major approach of data normalization [24] includes z-score normalization, min-max normalization, and normalization by decimal scaling.

Min -Mean normalization

Nevertheless, min-max normalization takes min and max values for normalization. Taking max values all the time not yields accurate normalization outcomes for resolving this issue rather than utilizing maximum value mean value is calculated in this research.It carries out a linear transformation on the original RA gene expression profiles and plots a value dataset (RA_D) of G to $RA_{D'}$ in the range$[New_{min}(G), New_{mean}(G)]$.It is computed by the subsequent formula:

$$RA_{D'} = \frac{[RA_D - \min(G)] * [New_{mean}(G) - New_{min}(G)]}{[mean(G) - \min(G)]} \quad (1)$$

Here mean (G) denotes the mean value of the gene (G) and min (G) denotes the minimum value of gene. Here min-mean normalization maps a value of Rheumatoid Arthritis (RA) dataset RAD of gene G to $RA_{D'}$ in the range [0,1], therefore $New_{min}(G) = 0$ and $New_{mean}(G) = 1$ are initialized. Below indicates the formulae of min-max normalization.

$$RA_{D'} = \frac{[RA_D - \min(G)]}{[mean(G) - \min(G)]} \quad (2)$$

The mean and standard deviation $std(G)$ is discussed as follows

$$= \sqrt{\frac{\sum_{i=1}^{N}(G_i - \bar{G})^2}{N}} \quad (3)$$

$$mean(G) = \frac{G_1 + G_2 + \cdots . + G_N}{N} = \frac{1}{N}\sum_{i=1}^{N} G_i \quad (4)$$

Z-Score normalization

It is known as zero-mean normalization in which the values for a gene G are normalized dependent upon the mean mean(G) and standard deviation of gene std(G). A dataset (RA_D) of G is normalized to $RA_{D'}$by the subsequent formula:

$$RA_{D'} = \frac{[RA_D - \text{mean}(G)]}{[std(G)]} \quad (5)$$

Normalization by Decimal Scaling

It normalizes by moving the decimal point of value of gene G. The amount of decimal points moved based upon the absolute value of gene G. A value (RA_D) of gene G is normalized to $RA_{D'}$ by the subsequent formula:

$$RA_{D'} = \frac{RA_D}{10^m} \quad (6)$$

Here m is known as the least integer and Max ($|RA_{D'}|$).

3.3. Gene selection processes

Gene selection is the process of preprocessing the input dataset with the aim of evaluating the available attributes with the

intension that just the genes associated data are maintained and inappropriate gene data are eliminated. Gene selection is beneficial while the dataset dimensionality is huge.

These Gene selection techniques are proposed for two reasons:

1. Decrease the search space by eliminating irrelevant variables
2. To increase the predictive power of the classifiers in supervised learning.

For these reasons the well-known gene selection approaches (IG, CHI, CGS, and Consistency Based Subset Evaluation (CS)) are proposed in this section.

a. Information Gain (IG)

Information Gain is commonly utilized for assessing the goodness of a gene [25] for classification of RA samples. Normally, IG is calculated based on the variance among the class entropy and the conditional entropy in the preprocessed gene data.

$$I(Cl, G) = H(Cl) - H(Cl|G) \tag{7}$$

Here cl is known as the gene class variable, G is called the gene value and $H(Cl)$denotes the entropy measure of information and the uncertainty of a random variable. Actually, provided a gene samples $GS = \{(x_1, x_2, \ldots, x_k, c)\}_i, i = 1,2,..n$, here xk is known as the value of the k^{th} gene and$k$' gene, c is called the subsequent class label be capable of calculating the IG of the a^{th}gene by

$$I(Cl, G) = p(k)\sum_{c\epsilon Cl} p(c|k)logp(c|k) + p(\bar{k})\sum_{c\epsilon Cl} p(c|\bar{k})logp(c|\bar{k}) - \sum_{c\epsilon Cl} p(c)logp(c) \tag{8}$$

Genes having greater IG score are ranked greater than genes having fewer score. The greater score of genes chosen for classification.

b. Chi-squared (CHI)

Chi-squared [26] is defined as the degree of independence amongthe pair of gene data. The higher CHI score of a gene is the most autonomous gene from the class variable. With the aim of calculating the CHI score, consider *N*as the total size of the training set, Q as the amount of time gene a and class *c*happen together, P as the amount of time class *c*happens derived of gene k, R is the amount of time gene k happens deprived of class c, S is the amount of time neither knor*c*happens. The CHI score is calculatedby

$$CHI(k, c) = \frac{N \times (QS - RQ)}{(Q + P) \times (R + S) \times (Q + R) \times (P + S)} \tag{9}$$

Especially, CHI assesses the value of gene data by calculating the Chi-squared statistic value regarding the class. The greater value of gene is chosen for classification.

c. Correlation Based Gene Selection (CGS)

CGS is a simple filter algorithm, which ranks gene subsets and identifies the value of gene or subset of genes in keeping with a correlation based heuristic evaluation function. The reason of CGS is to identify subsets, which encompass genes, which are extremely correlated with the class and uncorrelated with one another. The respite of genes must be eliminated. Redundant genes must be misplaced since they would be highly correlated with one or more of the residual genes. The approval of a gene would be based upon the degree to which it predicted classes in regions of the instance space not previously identified by other gene s. CGS $Merit_s$ gene subset evaluation function is depicted along these lines [27]

$$Merit_s = \frac{kr_{cf}}{\sqrt{k + (k+1)r_{ff}}} \tag{10}$$

here Merits is known as the heuristic "merit" of a gene subset s comprising k gene s, r_{cf} is called the mean gene -class correlation $g \in s$), and r_{ff} is known as the average gene -gene inter-correlation. This equation is, actually, Pearson's correlation, in which each and every variable is standardized. The numerator could be supposed of as providing a sign of how predictive of the class a group of genes are; the denominator of how much redundancy is amongst them. The heuristic manages inappropriate genes since they are weak predictors of the class. Redundant attributes are distinguished in contradiction of as they would be highly correlated with one or more of the other genes [27].

d. Consistency Based Subset Evaluation (CS)

Consistency Based Subset Evaluation agrees the class consistency rate as the evaluation measure. The notion is to get a set of features, which split the real dataset into subsets, which comprise one class majority. Consistency metric is one among the known consistency based gene selection [28].

$$Consistencey_s = 1 - \frac{\sum_{j=0}^{k}|D_j| - |M_j|}{N} \tag{11}$$

here s is known as gene subset, k is called the amount of genes in s, D_j is known as the amount of incidences of the j^{th}gene value combination, $M_j$is called the cardinality of the majority class for the j^{th} genes value, and N is known as the amount of genes in the original dataset [28].

3.4. Designing DNCM model

A Neutrosophic Cognitive Map (NCM) is known as a neutrosophic directed graph where as a minimum of one edge is in indeterminacy signified by dotted lines as shown in figure In this research, it denotes the causal association among RA notions.

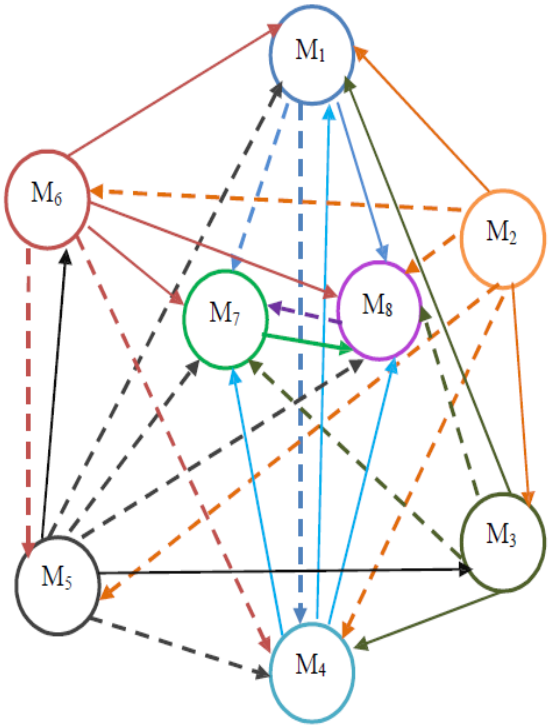

Figure 3: ANeutrosophic cognitive maps (NCMs) example

Neutrosophic logic

The Fuzzy Cognitive Maps (FCM)'s are handled by focusing the relation / non-relation among two nodes or ideas. however it is unsuccessful to manage the association among two conceptual nodes while the association is an uncertain one. Neutrosophic logic is the only tool well-known to us that handles the concepts of indeterminacy. It is a logic where in every proposition is guesstimated to contain the ratio of truth in a subset T, the ratio of indeterminacy in a subset I, and the ratio of falsity in a subset F, in which T, I, F are standard or non-standard real subsets. Consider$C_1, C_2, \ldots\ldots, C_n$represent n nodes and the nodes signify descriptive RA notions that could be features or behaviors of the

system. Moreover, every node is a neutrosophic vector from neutrosophic vector space V. Consequently a node $C_i$isdenoted by $(x_1, x_2, \ldots\ldots, x_k)$ in which x_k 's are 0or 1 or I (I is the indeterminate) and $x_k = 1$ signifies that the node C_k is in the greater phase of RA and $x_k = 0$denotes that the node is in the low phase of RA and $x_k = I$ denotes the nodes state is an intermediatephase of RA.

Consider$C_i$and $C_j$signify the two nodes as the RA notions of the NCM. The directed edge from $C_i$to $C_j$represent the causality of $C_i$on $C_j$known as connections. Each edge in the NCM is weighted with a number in the set {-1, 0, 1, I}. Consider$W_{ij}$is the weight of the directed edgeC_iC_j, $W_{ij} \in$ {-1, 0, 1, I}. W_{ij} = 0 if C_i doesn't contain any consequenceon C_j , W_{ij} = -1 when increase (or decrease) in$C_i$produces decrease (or increase) inC_j, $W_{ij} = -1$ when increase (or decrease) in $C_i$produces decrease (or increase) in C_j . W_{ij} = I when the relation or consequence of C_i on C_j is an indeterminate.

NCMs with edge weight from {-1, 0, 1, I} are known as simple NCMs. Consider$C_1, C_2, \ldots\ldots, C_n$are the nodes of a NCM. Consider the neutrosophic matrix $N(E)$isdescribed as N (E) = (W_{ij}) in whichW_{ij} is known as the weight of the directed edgeC_iC_j, in which $W_{ij} \in$ {- 1, 0, 1, I}. $N(E)$ Isknown as the neutrosophic adjacency matrix of the NCM.

Consider$C_1, C_2, \ldots\ldots, C_n$are the nodes of the NCM. Consider$A = (a_1, a_2, \ldots\ldots, a_n)$in which$a_i \in$ {-1, 0, 1, I}. A is known as the instantaneous state neutrosophic vector and it represents the on-off indeterminate state position of the node at a_n instant.

- $a_i = 0$ if a_i is off (low)
- $a_i = 1$ if a_i is on (high)
- $a_i = I$ if a_i is intermediate (medium) for $i = 1,2, \ldots\ldots, n$.

An NCM with cycles is said to contain a comment if there is a feedback in the NCM (i.e.) to say if the causal relations flow via a cycle in a revolutionary manner the NCM is known as a dynamical system. Consider C_1C_2 , C_2C_3 , C_3C_4 , … , $C_{n-1}C_n$ be cycle, if C_i is switched on and when the causality flow via the edges of a cycle and when it once more produces C_i, then the dynamical system goes round and round. This is factual for any nodec_i, for i = 1, 2,…, n. the equilibrium state for this dynamical system is known as the hidden pattern.

When the equilibrium state of a dynamical system is a unique state vector, it is known as a fixed point. Consider NCM is taken with$C_1, C_2, \ldots\ldots, C_n$ as nodes. Let's begin the dynamical system by switching onC_1. Presume that the NCM settles with C_1 and C_n on, that is to say the state vector go on as (1, 0,… 1) this neutrosophic state vector (1,0,…, 0, 1) is known as the fixed point. When the NCM settles with a neutrosophic state vector restating in the form$A_1 \to A_2 \to \cdots \to A_i \to A_1$, this equilibrium is known as a limit cycle of the NCM.

Dynamic NCM model

With the aim of expanding the ability of NCM, a Dynamic NCM (DNCM) model is proposedwhich is able to reflect dynamic conducts and designing nonlinear relationships in systems. For the adjacency matrix $N(E)$value to concept RAC_i, its relative position identifies its particular weight to other notions. As a resultW_{ij} is describedalong these lines:

$$w_{ij} = \begin{cases} 0 \in domain(low) \\ 1 \in domain(large) \\ -1 \in domain(no) \\ I \in domain\ (intermediate\ state) \end{cases} \quad (12)$$

W_{ij}consider three diverse values vigorously. Weight learning of NCM is equal to the optimization issue of the connection matrix that is resolved by means of the equation (12). NCM learning is concentrated on learning the adjacency matrix (a_i) and on the existing past raw RA data. The learning techniques for NCMs are concentrated on learning the adjacency matrix based on expert heuristic knowledge or on the existing past RA or on both of them. Evolutionary DNCM learning model to calculate adjacency matrices from past data that best fit the series of input state vectors. The learning objective of DNCM evolutionary learning is to produce ideal adjacency matrices for single NCMs modeling specific systems. Lastly, DNCMs is trained by hybrid learning methods

Step by step process dynamical system

1. Consider$C_1, C_2, \ldots\ldots, C_n$are the nodes of an NCM, with feedback, E is the related adjacency matrix.
2. Identify the hidden pattern whileC_1 is switched on while an input is provided as the vector A_1, the data must go through the neutrosophic matrix N(E), this is accomplished by multiplying $A_1$by the matrix N(E).
3. Weight learning by equation (12)
4. Consider $A_1N(E) = (a_1, a_2, \ldots\ldots, a_n)$ with the threshold operation, which is by substituting a_i by 1 when$a_i \geq k$ and a_i by 0 when$a_i < k$ (k – RA disease stage) and a_i by I whena_i is not an integer.
5. Bring up-to-date the resultant notion; the concept C_1 is encompassed in the updated vector by creating the first coordinate as 1 in the resulting vector.
6. Presume $A_1N(E) \to A_2$ take $A_2N(E)$ and do again the identical process.
7. This process is done again until obtain a limit cycle or a fixed point.

The learning objective is to change/bring up-to-date adjacency matrices from the preliminary heuristic expert's knowledge and past data at a multi-stage learning process. In this research, the researchers use Bat Algorithm (BA).

Bat Algorithm (BA)

The Bat Algorithm (BA) [29] is dependent upon ideal conduct of the echo-location ability of BAT. This presumes that *D*, the group of training NCM adjacency matrix and *F*, a candidate adjacency matrix are the inputs, and iteratively imposes the adjacency matrix. In every step, the adjacency matrix with high close to particle is selected by BA. Subsequently, the selected adjacency matrixes identified in all of the training occurrences are eliminated from *D* by marking it as chosen. With the smaller set of training dataset, the identical step is done again in the subsequent iteration. This process is repeated till, each and every training dataset is entire adjacency matrix are chosen. Fundamentally BA is a heuristic algorithm [29] whose idea is dependent upon the echo-location ability of micro bats, managing them on their scavenging conduct. In BA, the place of the bat denotes the best chosen adjacency matrix for the given classification issue. So, such a place identified by the i^{th} bat, is denoted as $Sa_i = (Sa_{i1}, \ldots, Sa_{iD})$ and the related fitness function is defined by fit_ithatrelates to the location quality.

i) Initiation of bats

Primarily the places of adjacency matrix are not well-known to the bats, consequently they will generate a distributed population *P* of *N* solutions in an arbitrary way, in which '*i*' represents the adjacency matrix count in *D* as the population. Inside the search space each solution is produced as provided (13).

$$Sa_{ij} = f_{min} + rand(0,1)(f_{max} - f_{min}) \quad (13)$$

here $i = 1, \ldots, N$ and $j = 1, \ldots, D$. Likewise in this equation f_{max}& f_{min}denotes the higher and lower bounds of the solutions for the adjacency matrix *I* and *j*, correspondingly while $rand(0,1)$ is utilized to produce the uniformly distributed value inside the range of [0,1].

ii) Generation of New Solutions

Dependent upon the information of the current position of the adjacency matrix and the place of the finest adjacency matrix, bats modify the selected minimum error of adjacency matrix solution. Bats traverse by regulating their flying directions by means of the finest experiences obtained on their own and other bat members' with the intention of identifying the finest position for the adjacency matrix selection issue. For all adjacency matrix position f_i, novel adjacency matrix is expressed along these lines:

$$v_i^t = v_i^{t-1} + (Sa_i^t - Sa^*)Sar_i \quad (14)$$
$$Sa_i^t = Sa_i^{t-1} + v_i^t \quad (15)$$

here 'i' denotes each adjacency matrix in the dataset, $i = 1, \dots, N$, and t represent the t^{th} iteration. Sa_i^t and v_i^t denotes the position and velocity components of the i^{th} adjacency matrix in the dataset, at the t^{th} iteration.

iii) Local Search

As soon as the novel adjacency matrix is selected, bat's random walk is invoked to carry out local search. While carrying out local search, when the pulse emission rate$r_i \in [0,1]$ of the i^{th} adjacency matrix is lesser compared to the random number, f_{old}^i is chosen from the dataset to generate a novel adjacency matrix position f_{new}^ialong these lines:

$$Sa_{new}^i = Saf_{old}^i + \varepsilon A^t \quad (16)$$

here f_{old}^i holds the solution selected from the present dataset and the arbitrary vector derived from a uniform distribution is denoted as ε. Likewise A^t in equation (16) represents the average loudness at the iteration (t), for each and every adjacency matrixes.

iv) Solutions, Loudness, and Pulse Emission Rate Updating

Presume that when a random number inclined to be greater compared to the loudness A^t and the condition $fit(Sa_{new}^i) > fit(Sa_i)$holds, after that agree$Sa_{new}^i$. Concurrently, if the pulse emission r_i is increased, the loudness A^t is decreased and expressed along these lines:

$$A_i^{t+1} = \alpha A_i^t \quad (17)$$
$$r_i^{t+1} = r_i^0(1 - e^{-\gamma t}) \quad (18)$$

α and γ are constants. A^0andr_i^0 represents the loudness and pulse emission rate. These values are the arbitrarily produced numbers in the range of [0,1] and [0, 1], correspondingly. The fundamental conduct of BA is as defined below (Algorithm 1).

Algorithm 1: Bat Algorithm(BA)

Input: Gene Samples GS, set of adjacency matrixes A ,

Output: Selected adjacency matrix // data driven built DNCM

1. Set iter=1 //where iter is the iteration parameter
2. Compute fitness function $fit_i(Sa)$

$$fit_i(sa) = Error = \frac{1}{M}\left(\sum_{i=1}^{N}\sum_{j=1}^{M}|fa_{ij} - \widehat{fa}_{ij}|\right) \quad (19)$$

Where N is the number of nodes, M is the number of profiles, $fa_{ij}$is the real value for the ithnode of the jthprofile, and $\widehat{fa}_{ij}$ is the predicted value for the ith node of the jth profile.

3. Initialize the adjacency matrix position and velocity of each adjacency matrix in the dataset
4. While iter $\leq$ iter_max do //where iter_max is the maximum number of iterations
   - 4.1. if rand (0,1)<A_i and $fit(Sa_{new}^i) > fit(Sa_i)$ Update the selected adjacency matrix solution, loudness and pulse emission rate by Eq. (16)- Eq. (18)
   - 4.2. end if // step 4.1
   - 4.3. Else Change the current node position and keep current adjacency matrix
5. end while // step 6
6. return best adjacency matrix

The proposed DNCM with BA 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 analysis, numerous arbitrary patients are generated and their severities of RA with gene profiles are computed by means of MATLAB tool.

4. Experimental results

Aspects accountable for radiographic harshness of Rheumatoid Arthritis (RA) in African-Americans are badly understood. Required to inspect genes whose expression in Peripheral Blood Mononuclear Cells (PBMCs) is related to radiographic harshness of RA. 20 control samples (from individuals not having RA) were matched up with 10 early severe, 10 early mild, 10 late mild, and 10 late severe RA samples. All samples were got from African-American individuals. With the intention of describing distinctive expression signatures in African American RA patients [30] with serious erosive disease, consider a gene expression analysis by utilizing samples of RNA from PBCs. The gene expression signature, which seems to associate with the amount of erosions at baseline and 36 months is an amount of transcriptional task related to RA severity, on the other hand not essentially disease development. The proposed DNCM-BA based unsupervised learning system and previous DFAFCM, FCM, FCM-PSO and DFCM classifiers are measured by means of the classification metrics and implemented by utilizing MATLAB simulation environment. According to predictive analytics, a table of confusion (known as a confusion matrix is displayed in table 1), is a table with two rows and two columns, which states the amount of false positives, true positives, false negatives, and true negatives. This lets more thorough examination compared to simple ratio of accurate presumptions (accuracy). Accurateness is not a consistent metric for the actual performance of a classifier, since it would produce correct results when the data set is unbalanced (to be exact, if the amount of samples in diverse classes differs significantly).

Table 1: Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	TP	FP
	Positive	FN	TN

TP is called the amount of accurate predictions that an instance is positive,

FN is called the amount of false predictions that an instance is negative,

FP is called the amount of false predictions that an instance positive, and

FN is called the amount of accurate predictions that an instance is negative.

Numerous standard terms are defined for the two class matrix:

The accurateness is the ratio of the entire amount of predictions, which were accurate. It is identified by means of the equation:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (20)$$

The Recall or True Positive Rate (TPR) is known as the ratio of positive cases, which were appropriately found, as calculated using the equation:

$$TPR = TP/(TP+FN) \quad (21)$$

The False Positive Rate (FPR) is known as the ratio of negatives cases, which were falsely categorized as positive, as computed by means of the following equation:

FPR= FP/ (FP+TN) (22)

The True Negative Rate (TNR) is known as the ratio of negatives cases, which were categorized appropriately, as computed by means of the equation

TNR= TN/ (TN+FP) (23)

The False Negative Rate (FNR) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

FNR= FN/ (FN+TP) (24)

Finally, Precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

Precision (P)= TP/ (TP+FP) (25)

The traditional F-measure or balanced F-score (F1 score) is known as the harmonic mean of precision and recall multiplying the constant of 2 scales the score to 1 while recall as well as precision are 1:

F-score= 2. (P.R)/(P+R) (26)

Table 2: 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	12

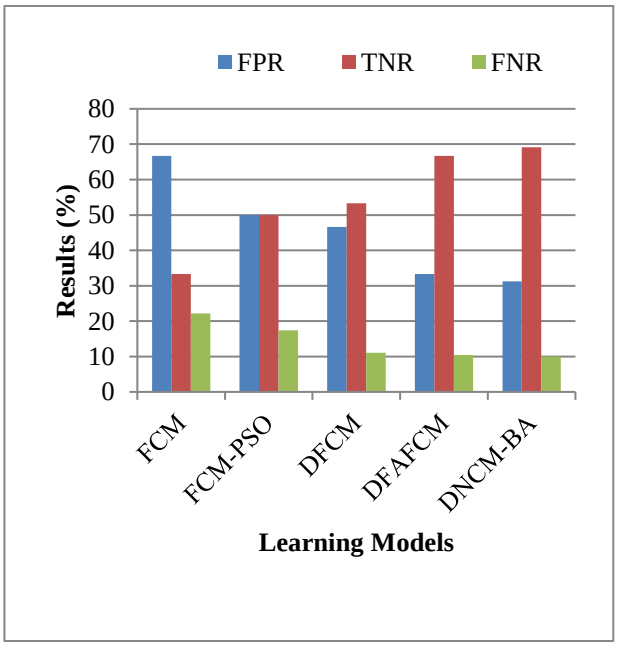

Figure 5: Learning models vs. results (TNR, FNR and FPR)

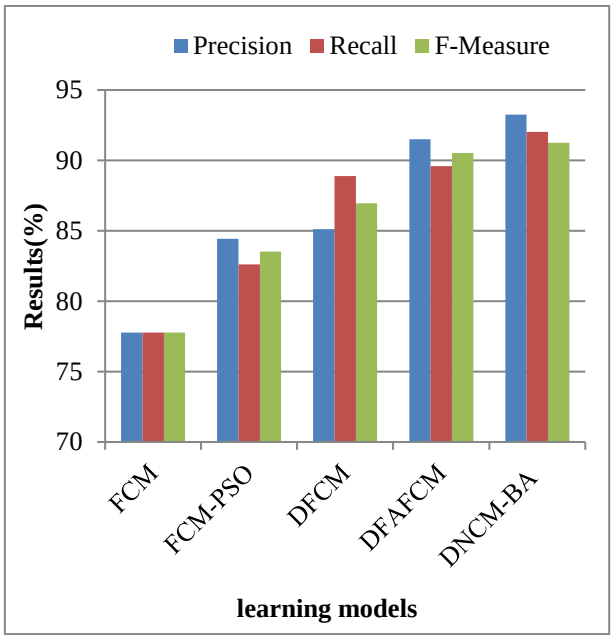

Figure 6: Learning models vs. outcomes (TNR, FNR and FPR)

Figure 5 indicates the performance evaluation results of five diverse classifiers for instance FCM, FCM-PSO [32], DFCM DFAFCM [31] and DNCM-BA (proposed) classifier in regard to TNR, FPR and FNR. It depicts that the proposed DNCM-BA classifier yields greater TNR detection percent and smaller FPR, FNR. The FCM classifier yields 66.67%, 33.33% and 22.22% for FPR, TNR and FNR correspondingly. The FCM-PSO classifier yields 50%, 50% and 17.39% for FPR, TNR and FNR correspondingly. The DFCM classifier yields 46.66%, 53.33% and 11.11% for FPR, TNR and FNR correspondingly. The DFAFCM classifier yields 33.33%, 66.77% and 10.42 % for FPR, TNR and FNR. The proposed DNCM-BA classifier yields 31.21%, 69.14% and 9.95% for FPR, TNR and FNR correspondingly. It confirms that the proposed DNCM-BA yields lower FNR outcomes of 9.95% that is 0.47%, 1.16%, 7.44% and 12.27% lower while matched up with DFAFCM, DFCM, FCM-PSO and FCM approaches correspondingly. It confirms that the proposed DNCM-BA yields greater TNR outcomes of 69.14% that is 13.34%, 16.67% and 33.34% greater while matched up with DFCM, FCM-PSO and FCM approaches correspondingly. It confirms that the proposed DNCM-BA identifies greater TNR when compared to other approaches.

Figure 6 indicates the performance assessment outcomes of four diverse classifiers for instance FCM, FCM-PSO, DFCM, DFAFCM and DNCM-BA (proposed) classifier in regard to recall, precision and f-measure. It indicates that the proposed DNCM-BA classifier yields greater f-measure, precision and recall. The FCM classifier yields 77.78%, 77.78% and 77.78% for precision, recall and f-measure correspondingly. The FCM-PSO classifier yields 84.44%, 82.61% and 83.52% for precision, recall and f-measure correspondingly. The DFCM classifier yields 85.11%, 88.89% and 86.96% for precision, recall and f-measure correspondingly. The DFAFCM classifier yields 91.49%, 89.58% and 90.53 % for precision, recall and f-measure. The proposed DNCM-BA classifier yields 91.49%, 89.58% and 90.53 % for precision, recall and f-measure correspondingly. It confirms that the proposed DNCM-BA yields greater f-measure outcomes of 90.53% that is 3.57%, 7.01% and 12.75% greater while matched up with DFCM, FCM-PSO and FCM approaches correspondingly. Proposed DNCM-BA yields greater recall outcomes of 89.58% that is 0.69%, 6.97% and 11.8% greater while matched up with DFCM, FCM-PSO and FCM approaches correspondingly. Proposed DNCM-BA yields greater precision outcomes of

91.49% that is 6.38%, 7.05% and 13.71% greater while matched up with DFCM, FCM-PSO and FCM approaches correspondingly.

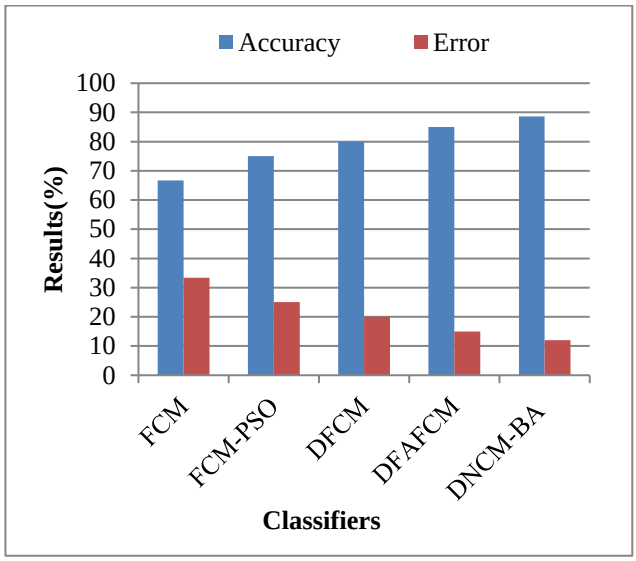

Figure 7: Learning models vs. outcomes (accuracy and error rate)

Figure 7 indicates the performance assessment outcomes of four diverse classifiers for instance FCM, FCM-PSO, DFCM and DNCM-BA (proposed) classifier in regard to accurateness and error rate. It indicates that the presented DNCM-BA classifier yields greater accurateness, and lower error value. The FCM classifier yields 66.67% and 33.33% for accurateness and error values correspondingly. The FCM-PSO classifier yields 75% and 25% for precision accurateness and error values correspondingly. The DFCM classifier yields 80% and 20% for accurateness and error values correspondingly. The proposed DNCM-BA classifier yields 85% and 15% for accurateness and error values correspondingly. It confirms that the proposed DNCM-BA yields greater accurateness outcomes of 85% that is 5%, 10% and 18.33% greater while matched up with DFCM, FCM-PSO and FCM approaches correspondingly.

5. Conclusion

In this research, a novel Dynamic Neutrosophic Cognitive Map (DNCM) with Bat Algorithm (BA) called DNCM-BA is utilized for finding out gene expression profiles, which differentiates patients with RA from strong control subjects. In this research, initially the data are preprocessed and after that gene is chosen by the IG and CHI techniques for increasing the classification accuracy. Next, the chosen gene expressions are sent to the learning process. In DNCM-BA learning model, genes are assessed whose expression in PBCs is related to radiographic harshness of RA. Twenty control samples (from individuals not having RA) were matched up with 10 early mild, 10 early severe, 10 late mild, and 10 late severe RA samples. All samples were got from African-American individuals. Performance assessment outcomes of four diverse classifiers for instance FCM, FCM-PSO, DFCM, DFAFCM and DNCM-BA (proposed) classifier in regard to accurateness and error rate. It confirms that the proposed DFAFCM yields greater accurateness outcomes of 88.63% that is 3.63%, 8.63%, 13.63% and 21.96% greater while matched up with the previous DFAFCM, DFCM, FCM-PSO and FCM approaches correspondingly. In the scope of future enhancement the prediction performance is improved by introducing hybrid feature selection techniques such as filter-wrapper-embedded algorithm.

References

[1] Gregersen PK, "Teasing apart the complex genetics of human autoimmunity: lessons from rheumatoid arthritis", *Clinical Immunology*, Vol.107, No.1,(2003), pp.1-9.

[2] Austen KF, Frank MM, Atkinson JP & Cantor HARVEY," Samter's immunologic diseases", *Philadelphia (PA): Lippincott Williams & Wilkins*, (2001).

[3] Weinblatt ME, Kremer JM, Bankhurst AD, Bulpitt KJ, Fleischmann RM, Fox RI, Jackson CG, Lange M & Burge DJ, "A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate", *New England Journal of Medicine*, Vol.340, No.4,(1999), pp.253-259.

[4] Zintzaras E, Dahabreh IJ, Giannouli S, Voulgarelis M & Moutsopoulos HM, "Infliximab and methotrexate in the treatment of rheumatoid arthritis: a systematic review and meta-analysis of dosage regimens", *Clinical therapeutics*, Vol.30, No.11, (2008), pp.1939-1955.

[5] Olivieri I, D'angelo S, Palazzi C & Padula A, "Advances in the management of psoriatic arthritis", *Nature Reviews Rheumatology*, Vol.10, No.9,(2014), pp.531-542.

[6] Staudt LM, "Gene expression profiling of lymphoid malignancies", *Annual review of medicine*, Vol.53, No.1,(2002), pp.303-318.

[7] Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, Shark KB, Grande WJ, Hughes KM, Kapur V & Gregersen PK, "Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus", *Proceedings of the National Academy of Sciences*, Vol.100, No.5,(2003), pp.2610-2615.

[8] Batliwalla FM, Baechler EC, Xiao X, Li W, Balasubramanian S, Khalili H, Damle A, Ortmann WA, Perrone A, Kantor AB & Gulko PS, "Peripheral blood gene expression profiling in rheumatoid arthritis", *Genes and immunity*, Vol.6, No.5,(2005), pp.388-397.

[9] Bomprezzi R, Ringner M, Kim S, Bittner ML, Khan J, Chen Y, Elkahloun A, Yu A, Bielekova B, Meltzer PS & Martin R, "Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease", H*uman molecular genetics*, Vol.12, No.17,(2003), pp.2191-2199.

[10] Dash M & Liu H, "Feature selection for classification", *Intelligent data analysis*, Vol.1, No.1-4,(1997), pp.131-156.

[11] Saeys Y, Inza I & Larrañaga P, "A review of feature selection techniques in bioinformatics", *bioinformatics*, Vol.23, No.19,(2007), pp.2507-2517.

[12] Zeng Z, Zhang H, Zhang R & Zhang Y, "A hybrid feature selection method based on rough conditional mutual information and naive Bayesian Classifier", *ISRN Applied Mathematics*, (2014).

[13] Uncu Ö & Türkşen IB, "A novel feature selection approach: combining gene wrappers and filters", *Information Sciences*, Vol.177, No.2,(2007), pp.449-466.

[14] Kohavi R & John GH, "Wrappers for feature subset selection", *Artificial intelligence*, Vol.97, No.1-2,(1997), pp.273-324.

[15] Yu L & Liu H, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", *Department of Computer Science & Engineering, Arizone State University, Tempe, AZ 85287-5406, USA*, (2003).

[16] Liu H & Yu L, "Toward integrating feature selection algorithms for classification and clustering", *IEEE Transactions on knowledge and data engineering*, Vol.17, No.4,(2005), pp.491-502.

[17] Wang H, Guo J, Jiang J, Wu W, Chang X, Zhou H, Li Z & Zhao, J, "New genes associated with rheumatoid arthritis identified by gene expression profiling", *International Journal of Immunogenetics*, Vol.44, No.3,(2017), pp.107-113.

[18] Chen YJ, Chang WA, Hsu YL, Chen CH & Kuo PL, "Deduction of Novel Genes Potentially Involved in Osteoblasts of Rheumatoid Arthritis Using Next-Generation Sequencing and Bioinformatic Approaches", *International Journal of Molecular Sciences*, Vol.18, No.11,(2017).

[19] Tchetina EV, ""High" and "Low" Gene Expression Signatures in Rheumatoid Arthritis: an Emerging Approach for Patient Stratification and Therapy Choice", *International Journal of Orthopaedics*, Vol.2, No.2,(2015), pp.219-226.

[20] Yoshida S, Arakawa F, Higuchi F, Ishibashi Y, Goto M, Sugita Y, Nomura Y, Niino D, Shimizu K, Aoki R & Hashikawa K, "Gene expression analysis of rheumatoid arthritis synovial lining regions by cDNA microarray combined with laser microdissection: up-regulation of inflammation-associated STAT1, IRF1, CXCL9,

CXCL10, and CCL5", *Scandinavian journal of rheumatology*,Vol.41, No.3,(2012), pp.170-179.
[21] Suzuki K, Yoshimoto K, Takeshita M, Kurasawa T & Takeuchi T, "THU0475 Identification of ATranscriptome-Wide Gene Expression Signature on Peripheral Blood from Patients with Systemic Lupus Erythematosus and Rheumatoid Arthritis by High-Throughput DNA Sequencing", *Annals of the Rheumatic Diseases*, Vol.73, (2014), pp.347-347.
[22] Giannopoulou EG, Elemento O & Ivashkiv LB, "Use of RNA sequencing to evaluate rheumatic disease patients", *Arthritis research & therapy*, Vol.17, No.1,(2015).
[23] Edwards III CK, Green JS, Volk HD, Schiff M, Kotzin BL, Mitsuya H, Kawaguchi T, Sakata KM, Cheronis J, Trollinger D & Bankaitis-Davis D, "Combined anti-tumor necrosis factor-α therapy and DMARD therapy in rheumatoid arthritis patients reduces inflammatory gene expression in whole blood compared to DMARD therapy alone", *Frontiers in immunology*, Vol.3, (2012), pp.1-10.
[24] Jain YK & Bhandare SK, "Min max normalization based data perturbation method for privacy protection", *International Journal of Computer & Communication Technology*, Vol.2, No.8,(2011), pp.45-50.
[25] Yu L & Liu H, "feature selection for high-dimensional data: A fast correlation-based filter solution", *Proceedings of the 20th international conference on machine learning*, (2003), pp.856-863.
[26] Sarkar SD & Goswami S, "Empirical study on filter based feature selection methods for text classification", *International Journal of Computer Applications*, Vol.81, No.6,(2013),pp.38-43.
[27] Karegowda AG, Manjunath AS & Jayaram MA, "Comparative study of attribute selection using gain ratio and correlation based feature selection", *International Journal of Information Technology and Knowledge Management*, Vol.2, No.2,(2010), pp.271-277.
[28] Yildirim P, "Filter based feature selection methods for prediction of risks in hepatitis disease", *International Journal of Machine Learning and Computing*, Vol.5, No.4,(2015), pp.258-263.
[29] Hasançebi O, Teke T & Pekcan O, "A bat-inspired algorithm for structural optimization", *Computers and Structures*, Vol.128, (2013), pp.77–90.
[30] https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-64707/files/
[31] Chithra B & Nedunchezhian R, "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 (JARDCS)*,15-Special Issue, (2017), pp. 171-186.
[32] Salmeron JL, Rahimi SA, Navali AM & Sadeghpour A, "Medical diagnosis of Rheumatoid Arthritis using data driven PSO–FCM with scarce datasets", *Neurocomputing*, Vol.232, (2017), pp.104-112.