ISSN 2278 – 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 23 - 33 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp04.pdf

DETECTION OF CIRCULATING TUMOR CELLS IN BREAST CANCER WITH HETEROGENEOUS CLUSTERING OPTIMIZATION ENSEMBLE FRAMEWORK

S.Mythili
Research Scholar, PG & Research Department of Computer Application, Hindusthan College of Arts &Science,
Coimbatore, India. poovanthikaa@gmail.com
Dr.A.V.Senthil Kumar
Director, PG & Research Department of Computer Applications, Hindusthan College of Arts & Science
Coimbatore, India. avsenthilkumar@yahoo.com

ABSTRACT: The study of Circulating Tumor Cells (CTCs) in peripheral blood resulting from tumor cell invasion and intravascular filtration highlights their crucial role concerning tumor aggressiveness and metastasis. In order to concerning tumor aggressiveness and metastasis problems, instead of direct CTC detection, in this paper focus on the novel Heterogeneous Clustering Optimization Ensemble Framework (HCOEF) is proposed for the identification of factors in peripheral blood (PB). Before performing clustering HCOEF methods, proposed work first stage, Decimal Scaling, Mix -Max and Z –Score Normalization schemas is applied to find missing values for gene samples. For preprocessed gene samples missing attributes data is imputated, and due to lack of indicative genes problem is solved by using Fuzzy Online sequential Ant colony Kernel Extreme Learning Machine (FOA-KELM) schema. The proposed FOA-KELM method the mean values for each gene feature is calculated and it is compared to ELM objective function to select and remove unimportant features. Fuzzy membership values of ELM are optimized by using Ant Colony Optimization (ACO). For selected gene features, HCOEF is proposed for the identification of CTC in BC. Proposed HCOEF combines the procedure of Hierarchical Randomized Firefly Clustering Algorithm (HRFCA), Hierarchical Differential Artificial Bee Clustering (HDABC) and Semi-Supervised Clustering (SSC) which classify the selected gene features into MS, NMS, MS and NMS. The cluster ensemble similarity measurement results are fused based on Weighted Quality (WQ), which in turn to improve classification results.

Keywords: Breast cancer (BC), Circulating Tumor Cells (CTCs), Weighted Quality (WQ), Semi-Supervised Clustering (SSC), Hierarchical Randomized Firefly Clustering Algorithm (HRFCA), Hierarchical Differential Artificial Bee Clustering (HDABC).

INTRODUCTION

Major progress has been achieved in the treatment of various cancers over the last decade. Nevertheless, metastatic cancer remains incurable. It is generally believed that distant metastases spread from the primary tumor through invasion into circulation and finally distant sites, where they may reinitiate growth, depending on the microenvironment. One of the surrogates of hematological spread are Disseminated Tumor Cells (DTC) as detected in bone marrow aspirations by Redding et al. in one of the first studies of its kind, using epithelial membrane antigen to identify DTC in bone marrow at the time of primary surgery in women with no overt metastases, and later by our group using a combination of epithelial cell surface antigens and cytokeratins to identify DTC in early stage breast carcinoma [1]. The prognostic significance of the presence of DTC in bone marrow aspirates has been established for several cancers, with most evidence related to breast cancer [2]. Owing to the invasive procedure, limited value of prognostic information and, as recently shown, low number of positive patients [3], bone marrow aspirations are not routinely performed. Many attempts have been made to optimize the detection of Circulating Tumor Cells (CTCs) from peripheral blood. Blood would be the ideal source because of the potential of serial cost–effectiveness and relative low invasiveness of the procedure. Since a blood draw is frequently performed in patients undergoing treatment, in the majority of cases only additional tubes need to be obtained from patients. Despite the progress in recent year, demonstrating not only the prognostic value, but also the use of CTCs in monitoring tumor responses to systemic therapy in breast and other cancers [4–5], the detection of CTCs is far from widely used in the clinical management of cancer patients.

However, recovery of tumor cells by this method is low, and enrichment is poor, thus the need for better enrichment methods became clear. The idea of using size for filtering CTCs from peripheral blood is an old one [6], however, it was not pursued until recently. Several techniques based on size-based separation of tumor cells are now either commercially available or under development [7-8]. Nowadays, a large number of high-dimensional gene expression datasets are obtained through the exploitation of molecular techniques, such as DNA microarrays. Gene expression profiling of CTCs might provide the opportunity to identify markers for diagnosis and prognosis in BC patients [9], toward better provision of personalized medicine [10]. Furthermore, exploring gene alterations in CTC profiles could give valuable information on the molecular mechanism of tumor cell metastasis. In general, several microarray studies on BC tissue samples demonstrate alterations in processes manifested in gene deformations. Similar gene alterations appear in the analyzed portion of PB [11]. In addition, Barbazan et al.

report that the spread of cancer relates to the detachment of malignant cells into blood [12] and Obermayer et al. [13] demonstrate that CTCs can be detected in single-cell level through specific genes (six gene panel) in PB. Particular microarray studies on PB that isolates specific CTC cells report that CTCs carry characteristics from the primary cause [13], but also convey information regarding the secondary (metastasis) tumor [14]. Thus, cancer-specific alterations can be identified in affected tissue areas, as well as in blood. Moreover, some specific alterations in cancer might be indicative of its ability to diffuse; such genes can indirectly predict the existence of CTCs without the need to detect and/or extract them [15].

To efficiently development of methods to reliably detect CTCs for BC curse dimensionality problem posses several challenges since it consist of lack of gene samples so it becomes very hard to find important genes and classify BC samples. To conquer all of these above mentioned problems gene samples, in this work study the procedure of novel gene selection method for CTC microarray technology. CTC identification the BC results are categorized into MS, NMS, MS and NMS it is also recommended for personalized medicine of patients. The FOA-KELM classification is proposed to select gene features is likely to reflect CTCs biology. For selected gene features then HCOEF is proposed to classify the gene samples into breast cancer tumor cells into three classes such as MS, NMS and combined MS and NMS. HCOEF method combines results of HRFCA, HDABC and SSC methods. WQ is proposed for the underlying similarity measurement among the cluster Ensemble Members such as HRFCA, HDABC and SSC which in turn can be highly classification results. It is used for monitoring and detection of CTCs for BC in genomic research area, classification results of HCOEF is better than previous conventional hierarchical clustering algorithms, since the proposed HCOEF ensemble clustering is performed and classification is performed based on semi supervised learning ,optimization methods. Experimentation results are performed to GSE29431 dataset samples for proposed HCOEF is evaluated using classification parameters. The remaining work of the paper is organized in the following manner: The related work comprises procedure of existing methods for identification of CTC in MS, NMS and BC; a major shortcoming of earlier methods is also discussed in Section II. Shortcoming of the present work is also finally mentioned in end of Section II; how it motivates the feature selection is also mentioned in Section II. The procedure of the proposed FOA-KELM for feature selection and hybrid HCOEF method for identification of CTC is described in Section III. The evaluation results of proposed HCOEF and existing hierarchical clustering methods is also discussed and experimented in Section IV. At end of the work finally concludes the results and major issues of the current work, several directions of the current work are also discussed in the Section V.

RELATED WORK

In [16] detection, enumeration and isolation of Circulating Tumour Cells (CTCs) have considerable potential to influence the clinical management of patients with breast cancer. There is, however, substantial variability in the rates of positive samples using existing detection techniques. This study was designed to directly compare three techniques for detecting CTCs in blood samples taken from 76 patients with Metastatic Breast Cancer (MBC) and from 20 healthy controls. In [17] completed a prospective, blinded study and evaluated CTC as a surrogate marker for treatment response and predictor of OS in patients with metastatic breast cancer. Patients were tested for CTCs at baseline, after 1 cycle of therapy, and at 12 weeks follow-up. The Overall Survival (OS) was defined as the time between the baseline blood draw and either death or last follow-up assessment. In [18] examined whether the five subtypes of breast cancer cells that have been defined by global gene expression profiling: normal-like, basal, HER2-positive, and luminal A and B were identified with CellSearch test. It did not recognize normal-like breast cancer cells, which in general have aggressive features.

In [19] conducted a prospective study that examined the correlation of CTCs with radiographic findings for disease progression. Shorter progression-fee survival was observed for patients with five or more CTCs at three to five weeks and at seven to nine weeks after the start of treatment. Potential limitations of the study include that the study included patients receiving various lines and types of therapy. In [20] conducted a prospective study to evaluate the accuracy of CTC in detecting localized and metastatic breast, colorectal, and prostate cancer. Blood samples were collected before the initiation of systemic therapy, and in patients with metastatic disease, samples were collected at least 3 weeks after surgery. Estimation of the area under the ROC curve (AUC) for breast cancer cases was 0.76 (95% CI, 0.72 to 0.81). The overall prognostic value of CTC was not reported.

The prognostic value of Circulating Tumor Cells (CTC) [21] detected in breast cancer patients is currently under debate. Different time points of blood collections and various CTC assays have been used in the past decades. The main outcomes analyzed were Overall Survival (OS) and Disease-Free Survival (DFS) in early-stage breast cancer patients, as well as Progression-Free Survival (PFS) and OS in metastatic breast cancer patients.

In a recent study, Peeters et al [22] also used the CellSearch pre-enrichment method followed by the DEPArray system for isolation and molecular characterization of single breast-tumor cell. However, one disadvantage of the

DEPArray is that there is approximately 40% cell-loss. Fabbri et al [23] isolated single CTCs from patients with metastatic colon cancer using as a pre-enrichment method a density gradient centrifugation, Onco-Quick (Greiner BioOne), followed by DEPArray. In summary, the existing CTC technologies rely on various properties of CTCs, with each having unique advantages and limitations. The inherent limitations of current CTC detection platforms should be considered when interpreting the literature about molecular properties of CTCs and their potential clinical applications.

HYBRID OPTIMIZED CLUSTERING ENSEMBLE FRAMEWORK FOR CTC IDENTIFICATION

Indeed, in this paper proposed a novel Heterogeneous Clustering Optimization Ensemble Framework (HCOEF). Proposed HCOEF methods the gene dataset samples is partitioned into n set of samples and clustering results combines the procedures of three clustering methods such as HRFCA, HDABC and SSC .In HRFCA, HDABC clustering methods distance based similarity is measured based on the procedure of optimization methods . Before performing clustering HCOEF method, in first stage the missing data imputation problem is solved by using the normalization methods .The proposed normalization methods missing data is imputated. Gene selection for preprocessed BC gene dataset samples, FOA-KELM is presented in this work. To select important gene features from BC gene dataset samples, mean value is calculated to each gene features and follows the procedure of KELM classification algorithm .In the proposed FOA-KELM method the fuzzy membership values is optimized by using Ant Colony Optimization (ACO) .Finally HCOEF method is applied to classify the selected gene feature samples as MS, NMS and MS and NMS. In HRFCA, HDABC clustering methods, first stage the gene samples from GSE29431 dataset is divisive, it split the original GSE29431 dataset recursively into small number of samples, and the second stage distance value is computed using DABC, RFA. Weighted Quality (WQ) is proposed for the underlying similarity measurement among the cluster Ensemble Members. The working procedure of the proposed schema is illustrated in Fig 1. In order to perform proposed HCOEF and FOA-KELM for Gene selection, let us consider each GSE29431 dataset samples matrix $\mathsf{d}(\mathsf{m},\mathsf{n})$ with m number of gene samples and n number of features from micro array data. The gene value for each feature is represented as range gr_{i}. Everyone of this dataset is registered with their own GEO format and downloaded individual platforms, it is preprocessed .Since the collected dataset samples are not constantly contains the complete dataset samples, so some preprocessing work is required to complete dataset samples and find missing values for gene samples. In this work normalization based methods have been used to find the missing value for GEO format dataset.

Preprocessing methods for missing data imputation

In data mining methods, the data imputation or missing data imputation problem is solved by using normalization. A missing attribute data of a GEO dataset is normalized through scaling its values from 0.0 to 1.0. In recent work several numbers of preprocessing methods is used among them data transformation methods, normalization methods [24] produces best missing gene attribute imputation method.

Scaling Normalization (SN): In this work missing gene data feature mgd problem is solved by using SN with decimal value. The number of decimal points value for gene features is moved based on higher gene feature sample value. The value of missed gene data feature m' is normalized based on the gene feature m and it is represented as $d(m,n)'$

Min Max Normalization (MMN) : In this work missing gene data feature mgd problem is solved by using MMN with decimal value. MMN is performed based on the linear transformation for missing data imputation . MMN maps a missed data gene value from $d(m,n)$ of m to $d(m,n)'$with the range values of $[new_{min(m)} = 0, new_{max(m)} = 1]$ is determined .

Z-Score Normalization (ZCN): In this work missing gene data feature mgd problem is solved by using ZSN with zero-mean for gene features. In the proposed ZSN schema, missing gene features data imputation is performed by calculation of mean and standard deviation for gene feature m.

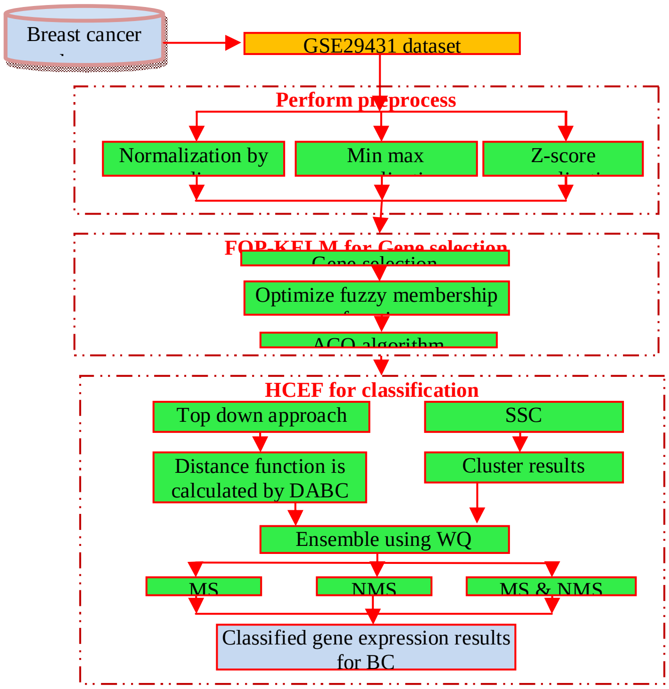

Fig 1: Overall illustration of the proposed work

FOA-KELM for Gene selection

This work presents a novel FOA-KELM method is applied for selection of gene from GSE29431 dataset samples. The proposed FOA-KELM comprises of two major steps: In the initial stage of the work, preprocessing is done based on normalization to finds missing values of the gene samples. In second stage of the work, KLEM method is applied to map the gene features from gene data matrix according to their linear KELM objective function [25]. With this FOA-KELM method, related GSE29431 dataset samples matrix in the same gene feature are gathered, determination significantly help improve BC detection results. The proposed KLEM method, fuzzy parameter (m) values are optimized using ACO algorithm. FOA-KELM method is proposed for gene selection from BC samples, which follows the procedure of Takagi–Sugeno–Kang (TSK) FISs [26]. The proposed FOA-KELM scheme, the fuzzy membership parameters $(\mathrm{c\ and\ a})$ are randomly generated which greatly reduces the gene selection results ,to shortcoming this problem $\mathrm{c\ and\ a}$ values are automatically generated using ACO , consequent parameters (β) are analytically determined. To reduce the complexity of the work , gene data samples is categorized into chunk by chunk is a necessity, where $\mathrm{d(m,n)'} \in \mathrm{GD_m} = (\mathrm{gf_1, \dots gf_n})$ $(\mathrm{GD_m},\ \mathrm{t_i})$, $\mathrm{GD_m}$ be the gene samples and $\mathrm{t_i} = [\mathrm{t_{i1}, t_{i2}, \dots, t_{im}}]^{\mathrm{T}} \in \mathrm{R_m}$, be the target gene feature selection results with L fuzzy rules is given as,

$$f_L(\mathrm{GD_m}) = \sum_{\mathrm{i=1}}^{\mathrm{L}} \mathrm{\beta_i G(GD_m, c_i, a_i)} = \mathrm{t_j}\,, \mathrm{j} = 1, ..\mathrm{N} \tag{1}$$

Consequent parameters of $f_L(\mathrm{GD_m})$is given by

$$\mathrm{\beta_i} = \mathrm{GD_{me}^T q_i} \tag{2}$$

where $\mathrm{GD_{me}^T}$ is the extended input gene data matrix vector $\mathrm{d(m,n)'}$ by appending the input gene data matrix vector $\mathrm{d(m,n)}$ and $\mathrm{q_i}$ is the parameter matrix for the i^{th} fuzzy rule is given by,

$$\mathrm{q_i} = \begin{bmatrix} q_{i1,0} & \cdots & q_{ip,0} \\ \vdots & \cdots & \vdots \\ q_{i1,o} & \cdots & q_{ip,o} \end{bmatrix} \tag{3}$$

The results of the TSK model is given in equation (4),

$$f_L(\mathrm{GD_m}) = \sum_{\mathrm{i=1}}^{\mathrm{L}} \mathrm{GD_{me}^T \beta_i G(GD_m, c_i, a_i)} = \mathrm{t_j}\,, \mathrm{j} = 1, ..\mathrm{N} \tag{4}$$

The equation is further extended to hidden matrix becomes,

$$HQ = \mathrm{T} \tag{5}$$

where H is the hidden matrix is given by,

$$H(c_1, \dots ., c_L, a_1, \dots a_L, \mathrm{GD_1}, \dots GD_m) = \left[\mathrm{GD_{me}^T(GD_1, c_1, a_1)}, \dots \mathrm{GD_{me}^T(GD_j, c_L, a_L)}\right] \tag{6}$$

$$Q = \begin{bmatrix} q_1 \\ \vdots \\ q_L \end{bmatrix} \tag{7}$$

When the hidden feature mapping function $h(x)$ is unknown, a kernel gene data matrix for ELM is given by:

$$H = \mathrm{h(GD_m, GD_k)} = \mathrm{EK(GD_m, GD_k)} \tag{8}$$

where $\mathrm{EK(GD_m, GD_k)}$ is a kernel function which may be any type of kernel function such as linear, and radial basis function. Proposed FOA-KELM scheme, the fuzzy membership parameters such as $(c\ and\ a)$ are randomly generated which greatly reduces the gene selection results; this is solved by using ACO. In FOA-KELM the fuzzy membership parameters such as $(c\ and\ a)$ is optimized based on ant colony optimization construction step , k number of ants is considered a probabilistic value is calculated to membership parameters $(c\ and\ a)$, named as random proportional rule, to decide selected parameters is optimized or not which is determined from city i to city j by ,

$$p_{ij}^k = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{l \in \mathcal{N}_i^k} [\tau_{ij}]^\alpha [\eta_{ij}]^\beta}\,, if\ j \in \mathcal{N}_i^k \tag{9}$$

where $\eta_{ij} = 1/d_{ij}$ is a heuristic optimized fuzzification parameter, α and β are two parameters to determine the optimized fuzzification parameters values in pheromone , $\mathcal{N}_i^k$ is the feasible fuzzification parameter for selected ant k at city i. By this optimized fuzzification parameter value in random proportional rule is associated pheromone based on the trail τ_{ij} is calculated by ,

$$\tau_{ij} = (1 - \rho)\tau_{ij} \tag{10}$$

$$\tau_{ij} \leftarrow \tau_{ij} + \sum_{k=1}^{m} \Delta\tau_{ij}^k \tag{11}$$

$$\Delta\tau_{ij}^k = \begin{cases} \dfrac{1}{C^k}\ arc(i,j) belongs\ to\ fitness \\ 0 \end{cases} \tag{12}$$

τ_{ij} is determined based on the $arc(i,j)$which belongs to fitness function ,highest clustering accuracy is considered as fitness function . From (12) better fuzzification parameters values is optimized for all values.

Heterogeneous Clustering Optimization Ensemble Framework (HCOEF)

To perform Heterogeneous Clustering Optimization Ensemble Framework (HCOEF) for selected gene features and it is partitioned into n set samples and it is represented as $\mathrm{ugfr_i, mgfr_i, hgfr_i}$,, i=1 to n from these values the clustering is formed and classified as samples. In this proposed HCOEF merge the results of three clustering methods. The cluster results are ensemble via the calculation of weight quality metric which combines the similarity measurement results of each cluster for CTC identification class especially for BC .When compare to single clustering methods the proposed HCOEF produces higher classification results for identification of CTC in BC with three classes . GSE29431 dataset selected gene features samples are also represented in multidimensional space $gf = (gf_1 \dots .\, gf_d)$. Divide the gene features samples as $gfr_i^u = [u(gfr_i)..y], gfr_i^m = [m(gfr_i)..y], gr_i^h = [h(gr_i)..y]$ into three ugr_i, mgr_i, hgr_i ranges and split as a hierarchical tree structure. The clustering results from above mentioned steps are considered as three classes $\langle gfr_i^u, gfr_i^m, gfr_i^h \rangle$ in hierarchical tree. The resulting cluster from the above mentioned steps is represented as $C = (c_1, \dots c_n)$, the clustering is formed based on the objective function which is specified in equation (14) and each one of the clusters belongs to individual classes. The sum of gene points in the cluster is specified as C_{gs}.The centroid value for each cluster C_{gs} is $C_{gs}^0 = \frac{gs}{N}$ .Consider Q be the Quadratic point which is the sum of each and every one of gene datapoints in the cluster.

$$\mathrm{Q_i} = \sum_{\mathrm{gfr} \in \mathrm{C}} \mathrm{gfr_i^2} \qquad (13)$$

The SSQ for cluster gene feature dataset samples as ,

$$\mathrm{SSQ_i}(\mathrm{C_{gs}}) = \sum_{\mathrm{gfr} \in \mathrm{C_{gs}}} \mathrm{dist_i^2}\,(\mathrm{gfr}, C_{gs}^0) \qquad (14)$$

$$= \sum_{gfr \in C} \sum_{i=1}^{n} (gfr_i - C_{gs}^0)^2$$

Hierarchical Random Firefly Clustering Algorithm (HRFCA): Firefly Algorithm (FA) is swarm intelligence based optimization algorithm is introduced and developed by Yang [19]. In this work Firefly Algorithm (FA) is used to calculate the distance between two genes datapoints in equation (15), global optimal distance value is found based on the flashing behavior of fireflies. In general, there are three basic rules are described to follow the procedure of the FA and described by Yang [19], is described as follows:

1) All cluster gene datapoints (fireflies) are unisex therefore that one cluster gene datapoints motivated to attract other cluster gene datapoints based on their sex;

2) Attractiveness is proportional to their brightness. Thus, for any two gene datapoints from the cluster, the fewer distance value of one gene point of the cluster will move towards better distance value of gene point, this form a cluster. If there is no better distance values are found for each gene data points, it will randomly select another cluster gene datapoints;

3) The brightness value of each firefly is determined based on the objective function defined in equation (14).

The attractiveness function β will be determined from objective function $\mathrm{dist_i^2}(\mathrm{gfr}, C_{gs}^0)$ and it is represented by:

$$\beta\left(\mathrm{dist_i^2}(\mathrm{gfr}, C_{gs}^0)\right) = \beta_0 e^{-\gamma \mathrm{dist_i^2}(\mathrm{gfr}, C_{gs}^0)} \qquad (15)$$

β_0 is represented as the attractiveness at $\mathrm{dist_i^2}(\mathrm{gfr}, C_{gs}^0) = 0$ and γ is the light absorption coefficient for each gene datapoint samples. The movement of a gene point datasamples gr , is attracted by another sum of gene points C_{gs} to form a cluster and it is determined by:

$$grff_i = grff_i + \beta_0 e^{-\gamma \mathrm{dist_i^2}(\mathrm{grf}, C_{gs}^0)}(grff_i - grff_j) + \propto \left(rand - \frac{1}{2}\right) \qquad (16)$$

$\propto \in [0,1], \gamma \in [0.01,100]$. The goal of proposed work is to show how the different randomized methods influence the results of the modified FA. In (16) $rand$ denotes the randomization function. Here the random value is generated between the intervals (0-1) from Uniform continuous distribution has the density function, as follows,

$$p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0\ otherwise \end{cases} \qquad (17)$$

Note that each possible value of the uniform distributed random variable is within optional interval $[a, b]$, on which the probability of each sub-interval is proportional to its length. If $a \le u < v \le b$ then the following relation holds:

$$p(u < x < v) = \begin{cases} \frac{v-u}{b-a} & u \le x \le v \\ 0\ otherwise \end{cases} \qquad (18)$$

Normally, the uniform distribution is obtained by a call to the random number generator . Note that the discrete variate functions always return a value of type unsigned which on most platforms means a random value from the interval[0, 2^{32} −1]. In order to obtain the random generated value within the interval [0, 1], the following mapping is used:

$$r = ((double)rand()/((double)(RAND\_MAX) + (double)(1))) \qquad (19)$$

where r is the generated random number, the function rand() is a call of the random number generator, and the RAND_MAX is the maximal number of the random value ($2^{32} - 1$).

Hierarchical Differential Artificial Bee Clustering (HDABC) : From the equation (15) the distance between two gene feature samples and sum of gene feature samples data matrix is calculated based on ABC. In ABC algorithm , the colony of artificial bees consists of three groups of bees: employed bees, onlookers and scouts for distance calculation between two gene feature samples. At the first step, the ABC randomly generates initial cluster gene feature samples data points $\mathrm{gfr_i}(\mathrm{i} = 1, 2, \ldots, \mathrm{SN})$as population, where SN denotes the size of cluster population. After initialization of the gene feature samples in the cluster, the population is subjected to repeated cycles, $\mathrm{C} = 1, 2, \ldots, \mathrm{MCN}$ until best distance calculation. Provided that the nectar amount of selected gene feature data point distance is smaller that of the previous one, the employee bee memorizes the new distance function and select gene feature samples as cluster and forgets the old one. Otherwise kept previous distance value in her memory. The detailed description of ABC is specified and discussed in [27]. This crucial drawback of ABC is limits their applications to offline problems with little or no real-time constraints. It starts with population of randomly distributed solutions . To solve these problem Differential Evaluation (DE) operations such as mutation, and crossover with simple arithmetic operator, which is based on differential vector between two randomly chosen parameter vectors to evolve the population. The basic idea behind DE is a scheme for generating trail parameter vectors. Mutation and crossover are used to generate new generations of artificial bee vectors (trial vectors), and selection then after determines which of the vectors have to be survive for the next generation.

Mutation : Mutation is a process in which DE generates a *donor* vector $\vec{V}_{i,G}$ corresponding to each population individual (member) or *target* vector $\vec{X}_{i,G}$ in the current generation, after initialization of population. The different strategies are distinguished by the following notation: DE/*x*/*y*/*z*, where DE stands for differential evolution, *x* represents a string denoting the vector to be perturbed, *y* is the number of difference vectors considered for perturbation of *x*, and *z* denotes the recombination strategy which is used to create the trail vector. For each target vector $\vec{X}_{i,G} = [x_{1,i,G}, x_{2,i,G}, \ldots x_{D,i,G}]$a mutant vector $V_{i,G}$ is generated according to:

$$"DE/rand/1{:}\, \vec{V}_{i,G} = \vec{X}_{r_1^i,G} + F.(\vec{X}_{r_2^i,G} - \vec{X}_{r_3^i,G}) \qquad (20)$$

The indices r_1^i, r_2^i, r_3^i are mutually exclusive integers randomly chosen from the range [1, *NP*], and all different from the base index *i*. The scaling factor *F* also called mutation factor between [0, 2] controls the amplification of the differential variation .

Crossover: To increase the potential diversity of the population, a crossover operation comes into play after generating the donor vector through mutation. The donor vector exchanges its components with the target vector *_Xi_G* under this operation to form the *trail* vector $\vec{U}_{i,G} = [u_{1,i,G}, u_{2,i,G}, \ldots u_{D,i,G}]$ This discrete recombination is adopted by DE using the following scheme.

$$u_{j,i,G} = \begin{cases} v_{j,i,G} & if\ (rand_{i,j}(0,1) \le Cr\ \ or\ j = j_{rand} \\ & x_{j,i,G}\ otherwise\ \ j = 1,.D \end{cases} \qquad (21)$$

where $rand_{i,j}(0,1) \in [0,1]$ is a uniformly distributed random number, which is called a new for each *j*th component of the *i*th parameter vector. $J_{rand} \in 1, 2, \ldots, D$ is a randomly chosen index, which ensures that $\vec{U}_{i,G}$ gets at least one component form $\vec{V}_{i,G}$ Here *CR* is a crossover constant which controls the recombination and lies between [0, 1]. The clustered from both algorithm points is grouped based on the following function,

$$\sum_{gfr \in C} gfr_i = C^0_{gfs}.N \qquad (22)$$

The working procedure of the proposed in top-down splitting in detail it is described as follows. At each iteration of the clustering process two basic steps are followed to complete the clustering process for all selected gene features from GSE29431 dataset samples and it is described as follows:

A) Choose anyone of the cluster gene feature samples data points $\mathrm{C_{gfs}}$ with largest SSQ value, and then

B) The separation of those selected cluster gene data sample points C_{gfs} based on the overall SSQ reduction, is represented as ΔSSQ.

Repeat these above mentioned two steps until the completion of selected gene features from GSE29431 dataset samples, where ΔSSQ is higher than the average value of SSQ. Load entire selected features GSE29431 dataset samples into the root node of Hierarchical Tree (HT) structure describe the function $\mathrm{Inttree}$. Once the completion of the load process of HT then starts the procedure of top down approach. From these steps clustering is formed and it is represented as function of $\mathrm{initclus}$ and the initial SSQ is determined for each cluster. The function $\mathrm{compavg\Delta SSQ}$ averages the real value of SSQ for all selected gene features from GSE29431 dataset. The function $\mathrm{compwegavg\Delta SSQ}$ is useful to the cluster $\mathrm{C_{gfs}}$ .The $\mathrm{weg\Delta SSQ}$ is determined based on the average value of SSQ attained through splitting $\mathrm{C_{gfs}}$ and reassignment of the cluster gene datapoint samples based on this splitting point based on these $\mathrm{avg\Delta SSQ}$ function .The working procedure of the hierarchical tree structure with FA is specified in detail [28].

Semi Supervised Clustering (SSC): In this paper proposes a Semi Supervised Clustering (SSC) [29] clustering methods to classify the partitioned gene feature selected dataset samples into three classes as mentioned above . Briefly, describe the procedure of Semi Supervised Clustering (SSC) algorithm by initialization of known set of selected gene feature samples from feature selection algorithm the samples is denoted as $N = (N_1, \dots N_l)$ such $0 \leq l \leq c$,where c be the total number of gene samples classes. At each iteration of the clustering process classification results if found for selected gene features π (in line 3). To perform clustering process and classify the selected gene features samples into MS,NMS and MS and NMS based on the selection criteria function x^* (at line 4) .For selected clustered gene feature samples $x^*$is then applied to queried user gene dataset samples beside each existing selected gene feature samples $N = (N_1, \dots N_l)$ to identify classification results and which is updated in (lines 5-12) based on the determined probability selected gene feature samples for best cluster N^*. This Semi Supervised Clustering (SSC) process for gene classification repeated until it meets the maximum number of iteration.

Algorithm 1: SSC for clustering

Input : Known set of selected gene feature samples from feature selection algorithm the samples is denoted as $N = (N_1, \dots N_l)$ such $0 \leq l \leq c$ with maximum number of iterations Max_{num}

Output : A clustering of $N = (N_1, \dots N_l)$ into c number of clusters with three classes

```
1.initialization $c = \emptyset$ ,$N = (N_1, \dots N_l)$ , $t = 0$
2. Repeat
3.$\pi = SSC(N, C)$
4.$x^* = Mostinformative(N, \pi, c)$
5. For each $N_i \in N$ in decreasing order for selected gene feature of $p(x^* \in N_i)$
6.do
7. For user selected gene feature point $x^*$ against $N_i$
8. t++;
9.Update cluster based on $x^*$
10. if $(x^*, N_i, MI)$ then $N_i = N_i \cup \{x^*\}$ break;
11. end for
12. else
13 then l++;$N_l = \{x^*\}$
14. Until $t > T$
15. Return clustering results
```

More properly, approximation the probability for selected gene feature samples x instance belonging to neighborhood gene feature samples N_i as,

$$p(x \in N_i) = \frac{\frac{1}{|N_i|} \sum_{x_j \in N_i} M(x_i, x_j)}{\sum_{p=1}^{l} \frac{1}{|N_i|} \sum_{x_j \in N_p} M(x_i, x_j)} \tag{23}$$

where$|N_i|$ is denoted as the total number of instances for selected gene features samples in neighborhood N_i, and l is the total number of presented gene feature samples , $M(x_i, x_j)$ similarity measurement between the two data instance for selected gene feature samples N_i .The uncertainty value of the gene feature samples is determined by using the entropy function,

$$H(N|x) = -\sum_{i=1}^{l} p(x \in N_i) \log_2 p(x \in N_i) \tag{24}$$

The clustering result from various clustering methods results is combined into single cluster in heterogeneous clustering framework is performed based weighted quality function. Normalized Partition cut is determined for each clustering results in equation (25).

$$NP_{cut}(\pi_a, \pi_b, \pi_c) = \frac{Mincut(\pi_a, \pi_b, \pi_c)}{\sum vol\,(\pi_a, \pi_b, \pi_c)} \tag{25}$$

$NP_{cut}(\pi_a, \pi_b, \pi_c)$ be the Normalized Partition cut of every clustering results for gene features and determined via (π_a, π_b, π_c) Ensemble cluster members π_a, π_b, π_c is determined via WQ_{abc}^{cl}.

$$WQ_{abc}^{cl} = \frac{1}{n_{cl}} \sum_{i=1}^{p} NP_{cut} \text{ ,} p = 1 \text{ to } 3 \tag{26}$$

Where n_{cl} denotes the total number of the clusters .From the equation (26) cluster results are ensemble for selected gene features and grouped as three classes.

EXPERIMENTATION RESULTS

Experimentation results of CTC is mainly relying on the datasets; to perform this process collect various categories of datasets from Gene Expression Omnibus (GEO) database [30] which is publicly available as open access. The details of these categories of dataset collected from GEO are shown in Table I with their appropriate characteristics. Among them these datasets much of the dataset provide information for both normal and cancer breast tissues. Moreover, these dataset will be collected from various platforms, but all of them Agilent and Affymetrix Human Genome be the mostly regularly used platforms, while one of the dataset using Applied Biosystems (ABI) and another using Agendia human Discoverprint V1 custom platform. In this work the following gene features are selected. Query subset A versus B: This feature recognize gene expression profiles of interest through determination of average rank among two gene feature samples in experimental subsets. Subset effects: Gene Profiles are standard if they demonstrate important dissimilarity in ranks values among subsets. This feature recovers each and every one gene profiles through value to a precise investigational variable, e.g. ‘age’ or ‘strain’. Value distribution: Box and whisker plots designed for each gene Sample inside GEO, allow an indication of the distribution of values across a GEO. GEO BLAST: These boundaries permit users to investigate intended for GEO Profiles of interest rely on nucleotide progression relationship by means of BLAST. In addition, usual BLAST output as achieve by means of NCBI's BLAST boundary, show ‘E’ icon associations where suitable, linking straightforwardly to GEO Profiles expression information. Cluster heat maps. Pre-computed example and gene hierarchical cluster heat up maps are present designed for the majority of Clusters distance metrics is calculated based on the procedure of swarm intelligence algorithm and then hierarchical clustering is applied ,SSC is also performed clustering. Multiple cluster ensembles results are grouped via WQ straightforwardly to Entrez GEO-Profile records.

First independent dataset (GSE29431) is introduced by [31] which provide information of the microarray Gene samples regarding 65 of primary breast arcinomas and 22 samples from normal breast cancer tissue types for BC patients. Consider a metastatic status regarding breast cancer samples which include 35 tumor samples, among 18 of them belong to metastatic as well as 17 of them belongs to non metastatic. To validate the results of the proposed HFCA clustering and existing hierarchical clustering algorithm for breast cancer samples, 24 genes were extracted from 14 tumor samples is used for validation. To validate clustering results of proposed HFA and existing hierarchical clustering algorithm for GSE29431 dataset the following metrics such as Sen, Spe, Pr, FPR, FNR and CA have been used in this work. The individual classification parameters definition and results of the existing hierarchical clustering and proposed HFCA also specified and discussed in the following ways.

TABLE 1: BREAST CANCER DATASETS

GEO	Origin	Platform	Cancer samples	Healthy samples
GSE22820	Tissue	Agilient whole human Genome Microarray 4 x44 k G4112F	176	10
GSE19783	Tissue	Agilient whole human Genome Microarray 4 x44 k G4112F	113	2(0*)
GSE31364	Tissue	Agendia human Discoverprint V1 custom platform	72	0
GSE9574	Tissue	Affymetrix Human Genome U133A array	14(0*)	15
GSE18672	Tissue	Agilient whole human Genome Oligo Microarray 4 x44 k G4112A	64	79
GSE27562	PB	Affymetrix Human Genome U133A plus 2.0 array	57	31
	PB	ABI human genome survey microarray version	67	54
GSE15852	Tissue	Affymetrix Human Genome U133A array	43	0
GSE12763	Tissue	Affymetrix Human Genome U133A plus 2.0 array	30	0

Precision (Pr): Precision is defined as percentages of predicted class which belongs to positive class that were correct, as determined using the equation:

$$\text{Precision} = \frac{A}{A+C} \qquad (27)$$

Sensitivity (Sen): Sensitivity is defined as the percentage of predicted and actual class which belongs to positive cases that were correctly identified, as determined using the equation:

$$\text{Sensitivity(Sen)} = \frac{A}{A+B} \qquad (28)$$

=(Number of true positive assessment)/(Number of all positive assessment)

Specificity (Spec): Specificity is defined as the percentage of predicted and actual class which belongs to negative cases that were correctly identified, as determined using the equation,

$$\text{Specificity (Spec)}=\frac{D}{D+C} \qquad (29)$$

=(Number of true negative assessment)/(Number of all negative assessment)

Classification Accuracy (CA): Classification accuracy is defined as the percentage of the total amount of predictions which belongs to both positive and negative cases that were correctly identified, as determined using the equation:

$$\text{Classification accuracy (CA)}=\frac{A+D}{A+B+C+D} \qquad (30)$$

=(Number of correct assessments)/Number of all assessments)

False positive rate (FPR): FPR is the defined as the percentage of predicted and actual class which belongs to negatives cases that were incorrectly classified as positive class, as determined using the equation:

$$\text{False positive rate (FPR)}=\frac{C}{C+D} \qquad (31)$$

False Negative **Rate (FNR) :** FNR is the defined as the percentage of predicted and actual class which belongs to positive cases that were incorrectly classified as negative class, as determined using the equation:

$$\text{False Negative Rate (FNR)}=\frac{B}{A+B} \qquad (32)$$

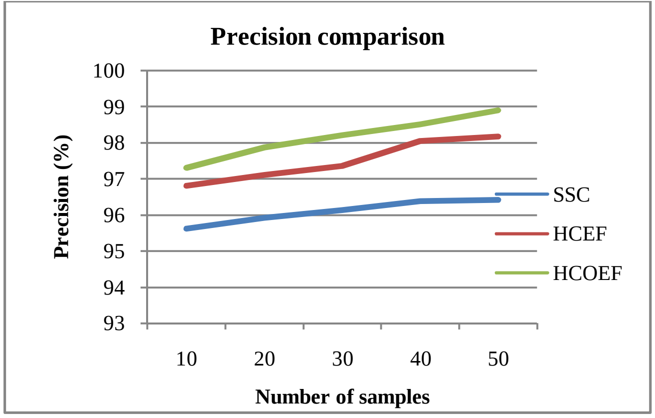

Fig 2: Precision comparison vs. methods

The precision results of proposed HCOEF and existing HCEF,SSC .Accuracy is represented as percentage of actual true positive results for GSE29431 dataset samples to identify CTC and detect CTC in BC which is illustrated in Fig 2. Since the proposed work normalization methods exactly finds the missed data values and instead of performing single clustering, proposed work HCOEF ensemble clustering is performed which combine the results of three clustering methods. Similarly precision results of proposed HCOEF clustering and Hierarchical clustering is defined as the percentages of predicted class which belongs to positive class, it shows that proposed HCOEF clustering schema have achieved higher than SSC clustering method is illustrated Fig.2, it is also applicable to all datasets where resultant will be changed based on the characteristics of the dataset. Since the proposed work gene features are selected using FOA-KELM methods which remove irrelevant features, cluster ensemble is performed instead of performing single clustering.

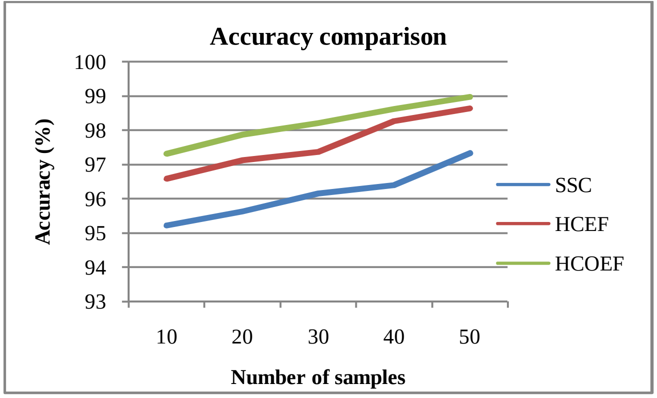

Fig 3: Accuracy comparison vs. methods

An accuracy of proposed HCOEF clustering IS increased when compared to HCEF, SSC clustering, so the test result shows that the contribution of the work is more accurate, regardless positive is illustrated in Fig.3.

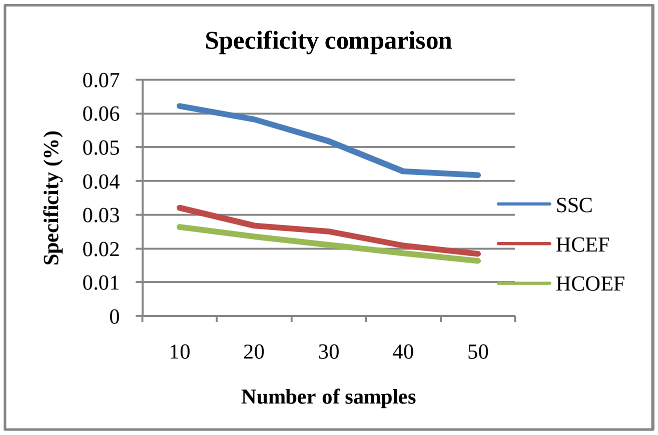

Fig 4: Specificity comparison vs. methods

Similarly specificity results of proposed HCOEF and Hierarchical clustering is defined as the percentage of predicted and actual class which belongs to negative cases, it shows that the proposed clustering methods have achieved 0.0163 %, which is 0.0254 % lesser than hierarchical clustering method is illustrated in Fig.4, it is also shown that the proposed system work performs well.

CONCLUSION AND FUTURE WORK

The isolation of CTCs though is difficult due to the small numbers of such disseminating cells in the PB. In this paper, describe a HCOEF approach that attempts to explore the field by combining microarray gene expression data originated from tissue and PB. HCOEF is proposed for the identification of CTC in BC. Proposed HCOEF combines the procedure of Hierarchical Randomized Firefly Clustering Algorithm (HRFCA), Hierarchical Differential Artificial Bee Clustering (HDABC) and Semi-Supervised Clustering (SSC) which classify the selected gene features into MS, NMS, MS and NMS. The results are indeed promising. The 24-gene signature contains some of the important genes that are commonly used for CTC identification, and therefore it makes sense to further consider the proposed approach for indirect assessment of the existence of CTCs. Proposed HCOEF attains better results when compare to conventional clustering methods, since the proposed work impact of randomized methods on the results of the RFA algorithms were verified. Proposed methods have therefore adopted the integration of the studies at the expression level using the optimization algorithm, which appears to yield the best results in various parameters. In future work, plan to examine the rest of genes in terms of their association with CTCs assessing their potential for direct identification of CTC cells. In the future, further experiments should be performed with the random sampling that exhibits the excellent results.

REFERENCES

1. Balic M. Disseminated tumor cells as biomarkers for breast cancer. Biomark Med. 2009; 3(3):215– 217.
2. Braun S, Vogl FD, Naume B, et al. A pooled analysis of bone marrow micrometastasis in breast cancer. N Engl J Med. 2005; 353(8):793–802.
3. Giuliano AE, Hawes D, Ballman KV, et al. Association of occult metastases in sentinel lymph nodes and bone marrow with survival among women with early-stage invasive breast cancer. JAMA. 2011; 306(4):385–393.
4. Bauernhofer T, Zenahlik S, Hofmann G, et al. Association of disease progression and poor overall survival with detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer. Oncol Rep. 2005; 13(2):179–184.
5. Miller MC, Doyle GV, Terstappen LW. Significance of circulating tumor cells detected by the cellsearch system in patients with metastatic breast colorectal and prostate cancer. J Oncol. 2010:617421.
6. Fleischer RL. Cancer filter deja vu. Science. 2007; 318(5858):1864.
7. Paterlini-Brechot P, Benali NL. Circulating tumor cells (CTC) detection: clinical impact and future directions. Cancer Lett. 2007; 253(2):180–204.
8. Vona G, Estepa L, Beroud C, et al. Impact of cytomorphological detection of circulating tumor cells in patients with liver cancer. Hepatology. 2004; 39(3):792–797.
9. L. Dirix, P. Van Dam, and P. Vermeulen, “Genomics and circulating tumor cells: Promising tools for choosing and monitoring adjuvant therapy in patients with early breast cancer?,” Curr. Opin. Oncol., vol. 17, no. 6, pp. 551–558, Nov. 2005.
10. S. Riethdorf and K. Pantel, “Advancing personalized cancer therapy by detection and characterization of circulating carcinoma cells,” Ann. New York Acad. Sci., vol. 1210, no. 1, pp. 66–77, Oct. 2010.
11. A. Balmain, J. Gray, and B. Ponder, “The genetics and genomics of cancer,” Nature Genetics, vol. 33, no. 3, pp. 238–244, Mar. 2003.
12. J. Barbaz´an, L. Alonso-Alconada, L. Muinelo-Romay, M. Vieito, A. Abalo, M. Alonso-Nocelo, S. Candamio, E. Gallardo, B. Fern´andez, I. Abdulkader, M. de Los A´ ngeles Casares, A. Go´mez-Tato, R. Lo´pez-L´opez, and M. Abal, “Molecular characterization of circulating tumor cells in human metastatic colorectal cancer,” PloS One, vol. 7, no. 7, p. e40476, 2012.
13. E. Obermayr, F. S. Cabo, M. K. Tea, C. Singer, M. Krainer, M. Fischer, J. Sehouli, A. Reinthaller,R.Horvat, G. Heinze, D. Tong, andR. Zeillinger, “Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients,” BMC Cancer, vol. 10, no. 1, p. 666, 2010.
14. T. J. Molloy, P. Roepman, B. Naume, and L. J. V. Veer, “A prognostic gene expression profile that predicts circulating tumor cell presence in breast cancer patients,” PloS One, vol. 7, no. 2, p. e32426, Feb. 2012.
15. M. Kandula, Ch. K. Kumar, K. Ravi Kanth, V. V. Laxmi Addala, S. Murthy, and Y. S. Ammi Raju, “Differences in gene expression profiles between human breast tissue and peripheral blood samples for breast cancer detection,” J. Cancer Sci. Ther., vol. 4, pp. 379–385, 2012
16. Van der Auwera I, Peeters D, Benoy IH, et al. Circulating tumour cell detection: a direct comparison between the CellSearch System, the AdnaTest and CK-19/mammaglobin RT-PCR in patients with metastatic breast cancer. Br J Cancer. 2010;102:276-284
17. Nakamura S, Yagata H, Ohno S, et al. Multi-center study evaluating circulating tumor cells as a surrogate for response to treatment and overall survival in metastatic breast cancer. Breast Cancer. 2010;17(3):199-204.
18. Sieuwerts AM, et al., “Anti-Epithelial Cell Adhesion Molecule Antibodies and the Detection of Circulating Normal-Like Breast Tumor Cells,” J Natl Cancer Inst. 2009; 101 (1): pp. 61–6
19. Liu MC, Shields PG, Warren RD, et al. Circulating tumor cells: a useful predictor of treatment efficacy in metastatic breast cancer. J Clin Oncol. 2009 Nov 1;27(31):5153-9.
20. Maestro Lm, Sastre J, Rafael SB, et al. Circulating tumor cells in solid tumor in metastatic and localized stages. Anticancer Res. 2009; 29(11): 4839-4844.
21. Zhang, L., Riethdorf, S., Wu, G., Wang, T., Yang, K., Peng, G., ... & Pantel, K. (2012). Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. Clinical Cancer Research, 18(20), 5701-5710.
22. Peeters DJ, De Laere B, Van den Eynden GG, Van Laere SJ, Rothe F, Ignatiadis M, Sieuwerts AM, Lambrechts D, Rutten A, van Dam PA, Pauwels P, Peeters M, Vermeulen PB,. Semiautomated isolation and molecular characterisation of single or highly purified tumour cells from Cell Search enriched blood samples using dielectrophoretic cell sorting. Br J Cancer. 2013;108:1358–1367.
23. Fabbri F, Carloni S, Zoli W, Ulivi P, Gallerani G, Fici P, Chiadini E, Passardi A, Frassineti GL, Ragazzini A, Amadori D. Detection and recovery of circulating colon cancer cells using a dielectrophoresis-based device: KRAS mutation status in pure CTCs. Cancer Lett.2013;335:225–231.

24. Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. International Journal of Computer & Communication Technology (IJCCT), 2(8), 45-50.
25. Chou, K. P., Prasad, M., Lin, Y. Y., Joshi, S., Lin, C. T., & Chang, J. Y. (2014). Takagi-Sugeno-Kang type collaborative fuzzy rule based system. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 315-320.
26. Liang, N.Y., Huang, G.B., Saratchandran, P. & Sundararajan, N. (2006). A fast and accurate online sequential learning algorithm for feed forward networks. IEEE Transaction on neural network, 17(6), 1411–1423.
27. Mythili, S., & Kumar, A. V. (2015, June). CTCHABC-hybrid online sequential fuzzy Extreme Kernel learning method for detection of Breast Cancer with hierarchical Artificial Bee. In Advance Computing Conference (IACC), 2015 IEEE International (pp. 343-348). IEEE.
28. Mythili, S. & Kumar, AVS. (2015). Discovery of Circulating Tumor Cells in Metastatic Breast Cancer and Nonmetastatic Cancer by Using Novel Hybrid Hierarchical Clustering Algorithm in Firefly Distance. Journal of Computer Technology & Applications, 6(1), 9-18
29. Xiong, S., Azimi, J., & Fern, X. Z. (2014). Active learning of constraints for semi-supervised clustering. IEEE Transactions on Knowledge and Data Engineering, 26(1), 43-54.
30. Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Tomashevsky, M. & Ron Edgar. (2007). "NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic acids research, D760-D765.
31. Lopez, F. J., Cuadros, M., Cano, C., Concha, A. & Blanco, A. (2012). Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. Medical Biological Engineering Computing, 50(9), 981–990.