IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

IDENTIFYING CIRCULATING TUMOR CELLS IN BREAST CANCER WITH DATA MINING ALGORITHMS BY USING MICROARRAY

Mythili Balu

DOI: 10.5281/zenodo.61445

ABSTRACT

Breast Cancer (BC) is an extremely diverse disease and extremely widespread among western women. Circulating tumor cells (CTCs) in Peripheral Blood (PB) is one of the most significant diagnostic factors for the cause of BC, the genomic study regarding CTCs detection in PB specifically for BC is limited because of lack of genes features for their identification and separation. As an alternative of direct CTC detection methods, in this study, we majorly focus on the Heterogeneous Swarm intelligent based Clustering Ensemble Framework (HSCEF) for the detection of distant factors in Peripheral Blood (PB). Proposed HSCEF combines the procedure of three clustering methods such as Hierarchical Levy Flights based Firefly Algorithm (HLFFA), Hierarchical Modified Artificial Bee Clustering (HMABC) and Semi-Supervised Clustering (SSC) which classify the selected gene features into Meta Static (MS), Non Meta Static (NMS), MS and NMS. In the proposed HSCEF, the similarity measurement results of various optimization methods are fused into single metric depending on Weighted Quality (WQ), which in turn to improve CTCs detection results in Peripheral Blood (PB). Publicly accessible breast cancer and PB microarray datasets is used in this work for experimentation of HSCEF procedure to the detection of several discriminant factors in Peripheral Blood (PB). Experimentation results is conducted to GSE29431 dataset samples and evaluated using the classification parameters like precision, specificity and classification accuracy.

KEYWORDS: Breast Cancer (BC), Circulating Tumor Cells (CTCs), Weighted Quality (WQ), Hierarchical Levy Flights based Firefly Algorithm (HLFFA), Hierarchical Modified Artificial Bee Clustering (HMABC).

INTRODUCTION

Circulating Tumor Cells (CTCs) are cancer cells initiated via from either a primary or metastatic tumor and circulate liberally in the Peripheral Blood (PB) [1]. It is identified that the detection and identification of CTC is a critical step in tumor metastasis [2]. In Breast Cancer (BC), CTCs be able to be detected in patients at initial stages or final stages of disease through explicit metastases [3-4]. Several numbers of the studies have been investigated in the literature for the identification of CTCs might help to predict the outcome of patients in together metastatic and non-metastatic BC patients [5]. Moreover, CTC count assessed on different time points throughout universal treatment is a consistent replacement marker of treatment response [6-8]. In recent work the some of the studies in the literature have suggested that therapies based methods for the identification of CTCs will improve treatment outcomes in patients [9-10]. Since CTCs are originated in distribution as a collectable fraction that is representative of the tumor,

they might present an ideal model to learning the biology of the tumor on different intervals before and for the period of treatment [11].

Currently, a huge number of high-dimensional gene expression datasets are obtained throughout the development of molecular techniques, such as DNA microarrays. Gene expression profiling of CTCs be able to provide new chance to recognize gene markers for identification and prediction in BC patients [12], in the direction of enhanced provision of personalized medicine [13]. In addition, exploring gene modification in CTC profiles might provide expensive information on the molecular categorization of tumor cell metastasis. In this research work obtain advantage of CTC microarray studies attained from the PB and tissue of BC patients, in order to devise a working hypothesis for the detection of a gene signature characterizing metastasis.

In general, several studies have been proposed in the recent work for the identification of CTC in BC tissue samples. Some of them are , Barbazan et al [14] proposed with the purpose of the spread of cancer regarding to the objectivity of malignant cells into blood [14] and Obermayer et al [15] experimented that the CTCs be able to be detected at single level during specific genes in PB. In specific some of the microarray studies on PB that studies the CTCs carry characteristics from the primary cause [16], however it also consider the information regarding the metastasis tumor [17], but these studies might not majorly focus on the CTC identification under gene signature. To solve this problem, Some of the studies in the earlier work, CTC identification is performed based on gene profile which is described as follows: Targeted studies have been described most important BP in BC [18], such as propagation, tumor incursion/metastasis, destruction of immune response, self-support in enlargement signals, and ER/HER2 signaling; however other key BP are expected to be added on the way to this list in the future Certainly, our 24 gene signature considers the information regarding both well-known and promising features of cancer, specifically the autophagy and the reprogramming of energy metabolism [19]. Furthermore, Shi et al [20] proposed a gene co-expression module examination reveal BP through the purpose of associated with BC progression. It consists of three major groups of modules, one of which (Group II) incorporated up-regulated modules such as cell series, RNA splicing, cellular part organization, and protein metabolic procedure with the purpose of associated to unrestrained cell propagation, a trademark of cancer . All these process have been applied to $C1 \cap C2 \cap C3$ intersection forming the CTC-related 24 gene signature and through enrichment examination to evaluate their relevance to tumor expansion, development, invasion, and metastasis [21].

Based on the previous observations, in this study, demonstrated that Heterogeneous Swarm intelligent based Clustering Ensemble Framework (HSCEF) process to $C1 \cap C2 \cap C3$ intersection forming the CTC-related 24 gene signature in metastasis [21] for identification of CTC. In the proposed HSCEF procedure, first stage aims to extract 24 gene signatures associated through pairwise discrimination among cell category and/or disease states. For instance, the evaluation of cancer and control tissue provides information regarding the discriminative factors of the primary disease. Next, the comparison among cancer blood and control PB is able to derive gene markers analytical of alterations appropriate to the pathology, associated to the CTC substance and in association through the primary and secondary disease. In these stage three different gene signatures is generated with the specific comparison which is specified as follows:
1) C1: comparison among cancer tissue and normal tissue in regulate to recognize genes expressed in primary cancer;
2) C2: comparison among PB of cancer patients and usual individuals, which recognize genes expressed in the blood of cancer patients; and
3) C3: comparison among cancer tissue and usual PB, which recognize genes over expressed in primary cancer and not in PB cells.

Summarizing this motivation, proficiently improvement of methods to consistently identify CTCs for BC with curse dimensionality problem posses several challenges, since it consist of lack of gene samples so it becomes very hard to find significant genes and classify BC samples. To overcome these problems, in this work HSCEF is applied for classification of the CTC microarray samples. The Fuzzy Online sequential Ant colony Kernel Extreme Learning Machine (FOA-KELM) method is proposed to select 24 gene signatures is likely to reflect CTCs biology. For selected gene features then HSCEF is proposed to classify the gene samples into breast cancer tumor cells into three classes such as MS, NMS, MS and NMS, it is also recommended for personalized medicine of patients. HSCEF method combines results of various clustering methods HLFFA, HMABC and SSC methods. WQ is proposed for the underlying similarity measurement among the cluster Ensemble Members such as HLFFA, HMABC and SSC which in turn higher classification results.

PROPOSED HETEROGENEOUS SWARM INTELLIGENT BASED CLUSTERING ENSEMBLE FRAMEWORK (HSCEF)

In this paper, we propose a HSCEF method for the identification of the gene signature in BC samples, samples is partitioned into n set of samples and results are combined from various clustering methods such as HLFFA, HMABC and SSC. In HLFFA, HMABC clustering methods distance based similarity is measured based on the procedure of optimization methods. Before performing clustering HSCEF method, in first stage the missing data imputation problem is solved by using the normalization methods. For preprocessed BC gene dataset samples, FOA-KELM is proposed for gene signature selection. To select important gene signature features, mean value is computed to each gene signature features and follows the procedure of FOA-KELM algorithm .In the proposed FOA-KELM method the fuzzy membership values of KELM is optimized by using Ant Colony Optimization (ACO) .Finally HSCEF method is applied to cluster the samples and classify those selected gene signature feature samples into three classes MS, NMS and MS and NMS. HSCEF method, first stage the selected gene signature features 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 LFFA and MABC. Weighted Quality (WQ) is proposed for the underlying similarity measurement among the various clustering methods.

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 with ranges from 0.0 to 1.0. In recent work several numbers of preprocessing methods is used applied to solve missing data imputation problems among them, normalization methods [22] produces best missing gene attribute replacement results, so in this work majorly consider normalization method 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) : 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 the proposed ZSN schema, missing gene features data imputation is performed by calculation of mean and standard deviation for gene feature m.

The working procedure of the proposed schema is illustrated in Figure 1. In order to perform proposed HCOEF and FOA-KELM for Gene selection, let us consider each GSE29431 dataset samples matrix $d(m,n)$ with m number of gene samples and n number of features from micro array data. The gene signature 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 signature feature samples. In this work normalization based methods have been used to find the missing value for GEO format dataset.

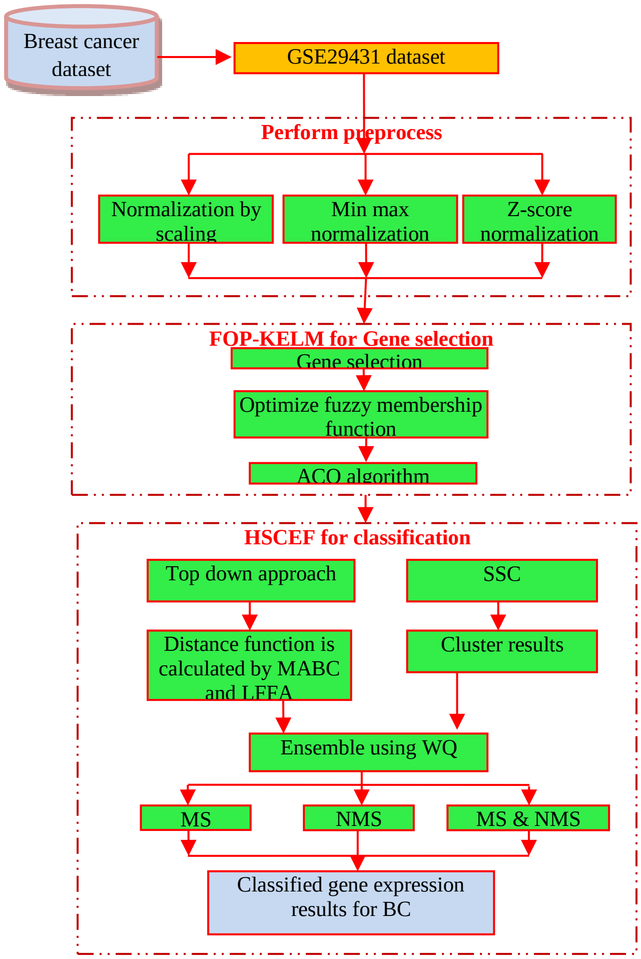

Figure 1: Overall illustration of the proposed work

FOA-KELM FOR GENE signature SELECTION: This work presents a novel FOA-KELM method for selection with two major steps: In the initial stage of the work, preprocessing is done based on normalization method to solve missing data gene imputation problem. In second stage of the work, KLEM method is applied to map the gene signature features from gene data matrix according to their linear KELM objective function [23]. FOA-KELM method is proposed for gene selection from BC samples, which follows the procedure of Takagi–Sugeno–Kang (TSK) FISs [24]. In general KELM scheme, the fuzzy membership parameters (c and a) are randomly generated which affects the gene signature selection results, to shortcoming this problem 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 $\mathrm{d(m,n)' \in GD_m = (gf_1, \dots gf_n)\ (GD_m, t_i), GD_m}$ be the gene samples and $\mathrm{t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in R_m}$, be the target gene feature selection results with L fuzzy rules is given as,

$$\mathrm{f_L(GD_m) = \sum_{i=1}^{L} \beta_i G(GD_m, c_i, a_i) = t_j\,, j = 1, .. N, \beta_i = GD_{me}^T q_i, q_i = \begin{bmatrix} q_{i1,0} & \cdots & q_{ip,0} \\ \vdots & \cdots & \vdots \\ q_{i1,o} & \cdots & q_{ip,o} \end{bmatrix}} \qquad (1)$$

where $\mathrm{GD_{me}^T}$ is the extended input gene data matrix vector $\mathrm{d(m,n)'}$ by appending the input gene signature data matrix vector $\mathrm{d(m,n)}$ and $\mathrm{q_i}$ is the parameter matrix for the $\mathrm{i^{th}}$ fuzzy rule .The results of the TSK model is given in equation (2),

$$\mathrm{f_L(GD_m) = \sum_{i=1}^{L} GD_{me}^T \beta_i G(GD_m, c_i, a_i) = t_j\,, j = 1, .. N} \qquad (2)$$

The equation is further extended to hidden matrix becomes,

$$\mathrm{HQ} = \mathrm{T}, \mathrm{Q} = \begin{bmatrix} \mathrm{q}_1 \\ \vdots \\ \mathrm{q}_\mathrm{L} \end{bmatrix} \qquad (3)$$

$$\mathrm{H}(\mathrm{c}_1, \ldots., \mathrm{c}_\mathrm{L}, \mathrm{a}_1, \ldots \mathrm{a}_\mathrm{L}, \mathrm{GD}_1, \ldots \mathrm{GD}_\mathrm{m}) = \left[\mathrm{GD}_{\mathrm{me}}^{\mathrm{T}}(\mathrm{GD}_1, \mathrm{c}_1, \mathrm{a}_1), \ldots \mathrm{GD}_{\mathrm{me}}^{\mathrm{T}}\left(\mathrm{GD}_\mathrm{j}, \mathrm{c}_\mathrm{L}, \mathrm{a}_\mathrm{L}\right)\right],$$

where H is the hidden matrix is given in (3). When the hidden feature mapping function $\mathrm{h}(\mathrm{x})$ is unknown, a kernel gene data matrix for ELM is given by:

$$\mathrm{H} = \mathrm{h}(\mathrm{GD}_\mathrm{m}, \mathrm{GD}_\mathrm{k}) = \mathrm{EK}(\mathrm{GD}_\mathrm{m}, \mathrm{GD}_\mathrm{k}) \qquad (4)$$

where $EK(GD_m, GD_k)$ 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 $(c\ and\ a)$ is optimized using ACO in construction step with k number of ants .The probabilistic value is calculated to membership parameters $(c\ and\ a)$, named as random proportional rule, to decide selected ELM parameters is optimized or not from city i to city j and described as,

$$p_{ij}^k = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{l \in \mathcal{N}_i^k} [\tau_{ij}]^\alpha [\eta_{ij}]^\beta}\ , if\ j \in \mathcal{N}_i^k \qquad (5)$$

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 ,

$$\tau_{ij} = (1 - \rho)\tau_{ij} \qquad (6)$$

$$\tau_{ij} \leftarrow \tau_{ij} + \sum_{k=1}^{m} \Delta\, \tau_{ij}^k \qquad (7)$$

$$\Delta\tau_{ij}^k = \begin{cases} \dfrac{1}{C^k}\ arc(i,j) belongs\ to\ fitness \\ 0 \end{cases} \qquad (8)$$

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.

HSCEF

To perform HSCEF for selected gene signature features and it is partitioned into n set of gene signature feature samples and it is represented as $ugfr_i, mgfr_i, hgfr_i$,, i=1 to n from these values the clustering is formed and classified as three classes. In this proposed **HSCEF** merge the results of three clustering methods. The clustering results of various clustering results are ensemble via the calculation of weight quality metric for CTC identification in BC .When compare to single clustering methods the proposed **HSCEF** produces higher classification results for identification of CTC in BC with three classes. GSE29431 dataset selected gene signature features samples are also represented in multidimensional space $gf = (gf_1 \ldots. gf_d)$. Divide the gene signature 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 resulting cluster from the above mentioned steps is represented as $C = (c_1, \ldots c_n)$, the clustering is formed based on the objective function which is specified in equation (10) 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.

$$Q_i = \sum_{gfr \in C} gfr_i^2 \qquad (9)$$

The SSQ for cluster gene feature dataset samples as ,

$$SSQ_i(C_{gs}) = \sum_{gfr \in C_{gs}} dist_i^2\left(gfr, C_{gs}^0\right) = \sum_{gfr \in C} \sum_{i=1}^{n} (gfr_i - C_{gs}^0)^2 \qquad (10)$$

Hierarchical LFFA (HLFFA): Firefly Algorithm (FA) is swarm intelligence based optimization algorithm is introduced and developed by Yang [25]. In this work Firefly Algorithm (FA) is used to calculate the distance between two genes signature datapoints in equation (11), 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 [25], 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 (10).

Further increase the clustering results , randomness is added to FA in order to explore new distance gene signature data points by moving the towards the search space. The attractiveness function β will be determined from objective function $dist_i^2(gfr, C_{gs}^0)$ and it is represented by:

In the **LFFA** implementation, the actual form of attractiveness function $\beta(dist_i^2(gfr, C_{gs}^0))$ can be determined from objective function $dist_i^2(gfr, C_{gs}^0)$ with monotonically decreasing functions and it is represented the following generalized form,

$$\beta\left(dist_i^2(gfr, C_{gs}^0)\right) = \beta_0 e^{-\gamma dist_i^2(gfr, C_{gs}^0)^m}, m \geq 1 \qquad (11)$$

β_0 is represented as the attractiveness at $dist_i^2(gfr, C_{gs}^0) = 0$ and γ is the light absorption coefficient for each gene datapoint samples. For a fixed γ, the characteristic length becomes $\Gamma = \gamma^{-\frac{1}{m}} \rightarrow 1$ as m → ∞. Conversely, for a given length scale Γ in an optimization problem, the parameter γ can be used as a typical initial value. That is $\Gamma = \frac{1}{\Gamma^m}$. The distance between any two gene signature feature datapoints in the fireflies i and j at $grff_i$ and $grff_j$.

$$grff_i = grff_i + \beta_0 e^{-\gamma dist_i^2(grf, C_{gs}^0)}(grff_i - grff_j) + \propto sign\left(rand - \frac{1}{2}\right) \oplus Levy \qquad (12)$$

$\propto \in [0,1], \gamma \in [0.01,100]$, In (16) $rand$ denotes the randomization via Levy flights The product $\oplus$ means entrywise multiplications. The $\alpha sign\left[Rand - \frac{1}{2}\right]$where rand ∈ [0, 1] essentially provides a random sign or direction while the random step length is drawn from a Levy distribution,

$$Levy \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \qquad (13)$$

which has an infinite variance with an infinite mean. Here the steps of firefly motion are essentially a random walk process with a power-law step-length distribution with a heavy tail.

Modified Artificial Bee Clustering (MABC) : From the equation (10) 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 $gfr_i(i = 1, 2, \dots, SN)$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, $C = 1, 2, \dots, MCN$ 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 [26]. In general ABC algorithm has been shown to be fine at examination however poor at development [27-28]. To solve this problem new neighbor gene signature data points distance calculation solution is generated by various distance search equation chosen randomly from a candidate pool, which consists of various selection strategies. Global best guided ABC [27] described a new gene signature data points distance calculation solution equation to improve the utilization and to take advantage of the global best gene signature data points distance solution information as follows:

$$v_{ij} = grf_{i,j} + \Phi_{ij}\left(grf_{i,j} - grf_{k,j}\right) + \Psi_{i,j}\left(dist_j - grf_{i,j}\right) \qquad (14)$$

where is $dist_j$ the jth selected gene signature feature element of global best distance solution, $\Phi_{ij}$is a uniform random number and is suggested to be 1.5, $\Psi_{i,j}$ is a random number in the range [-1,1] . ABC/best/1 [28] is performed based on the differential evolution algorithm with modified gene signature data points distance search equation as follows:

$$v_{ij} = dist_{best,j} + \Phi_{ij}\left(dist_{r_1,j} - dist_{r_2,j}\right) \qquad (15)$$

where the indices r_1 and r_2 are mutually exclusive integers randomly chosen; $dist_{best,j}$ is the best distance solution in the current selected gene signature feature samples population and $\Phi_{ij}$is a random number in the range .

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,

$$p(x \in N_i) = \frac{\frac{1}{|N_i|}\sum_{x_j \in N_i} M(x_i, x_j)}{\sum_{p=1}^{l} \frac{1}{|N_i|}\sum_{x_j \in N_p} M(x_i, x_j)} \qquad (16)$$

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,

$$H(N|x) = -\sum_{i=1}^{l} p(x \in N_i)\, log_2\, p(x \in N_i) \tag{17}$$

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 (18).

$$NP_{cut}(\pi_a, \pi_b, \pi_c) = \frac{Mincut(\pi_a, \pi_b, \pi_c)}{\sum vol\,(\pi_a, \pi_b, \pi_c)} \tag{18}$$

$NP_{cut}(\pi_a, \pi_b, \pi_c)$ 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}.

$$WQ_{abc}^{cl} = \frac{1}{n_{cl}} \sum_{i=1}^{p} NP_{cut}\, , p = 1\; to\; 3 \tag{19}$$

Where n_{cl} denotes the total number of the clusters .From the equation (19) cluster results are ensemble for selected gene features and grouped as three classes. 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 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 . 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 $initclus$ and the initial SSQ is determined for each cluster. The function $compavg\Delta SSQ$ averages the real value of SSQ for all selected gene features from GSE29431 dataset. The function compwegavgΔSSQ is useful to the cluster $\mathrm{C_{gfs}}$ .The wegΔSSQ 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 avgΔSSQ function .The working procedure of the hierarchical tree structure with FA is specified in detail [24].

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. 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 for GSE29431 dataset the following metrics such as Sen, Spe, Pr, and CA have been used in this work.

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:

$$\text{Precision} = \frac{A}{A+C} \tag{20}$$

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,

$$\text{Specificity (Spec)} = \frac{D}{D+C} \tag{21}$$

=(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:

$$\text{Classification accuracy (CA)} = \frac{A+D}{A+B+C+D} \tag{22}$$

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

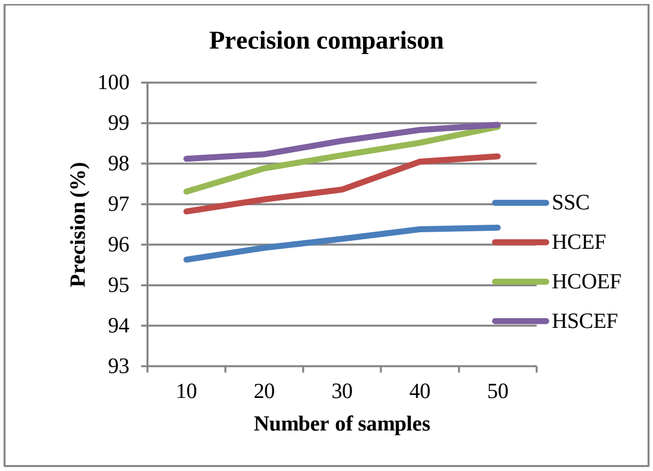

Figure 2: Precision comparison vs. methods

The precision results of proposed HSCEF and existing HCOEF HCEF,SSC to identify CTC and detect CTC in BC which is illustrated in Figure 2. Proposed HSCEF produces best clustering results since it perform clustering based on the ensemble clusters. Similarly the proposed work gene features are selected using FOA-KELM methods which remove irrelevant features, cluster ensemble is performed instead of performing single clustering.

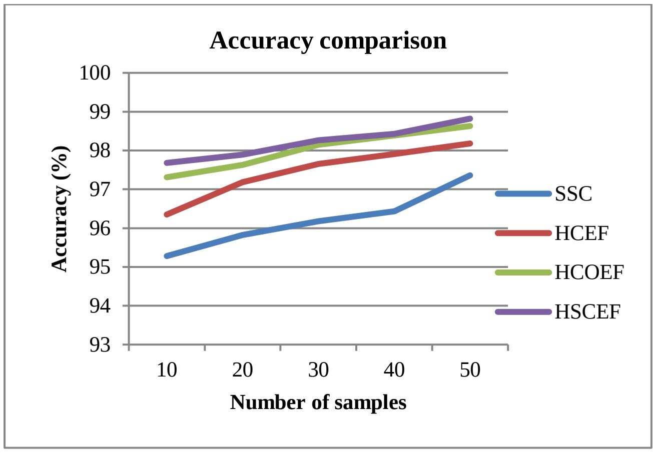

Figure 3: Accuracy comparison vs. methods

The accuracy results of proposed HSCEF and existing HCOEF HCEF, SSC to identify CTC and detect CTC in BC which is illustrated in Figure 3. Proposed HSCEF produces best clustering accuracy results, since it perform clustering based on the ensemble clusters. Similarly the proposed work gene features are selected using FOA-KELM methods which remove irrelevant features; cluster ensemble is performed instead of performing single clustering.

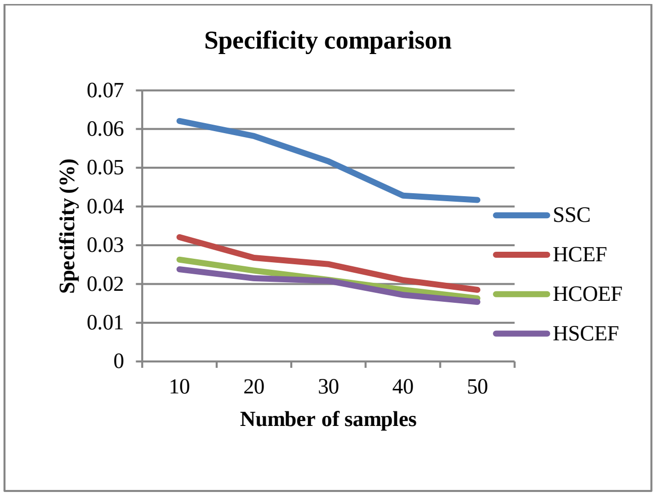

Figure 4: Specificity comparison vs. methods

Similarly specificity results of proposed HSCEF 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 0.01538 %, wheres the specificity results of the existing methods such as HCOEF, HCEF and SSC are 0.0163 %, 0.0185 % and 0.0417 % respectively as is illustrated in Figure.4, it is also shown that the proposed HSCEF system work performs well.

CONCLUSION AND FUTURE WORK

The detection and classification of CTCs is essential for the treatment of patients in metastatic epithelial cancers especially for BC. The separation of CTCs is difficult due to the minute numbers of cells in the PB. Several numbers of methods have been proposed in recent work to identify CTCs in BC, however in this paper work discuss the identification of the CTC in BC with PB for gene signatures markers. Describe a HSCEF approach with the purpose of attempts to discover the field through combining microarray gene expression data originated from tissue and PB. Proposed HSCEF combines the procedure of HLFFA, HMABC and SSC. Proposed HSCEF presently recommend new methodology to identify CTCs at a clinical trial, but encourage further work in the field of microarray technology improvement at metastatic process. Proposed HSCEF attains better results when compare to conventional clustering at the 24 gene expression signature level. In future work, plan to examine the identification of CTC on several cancer types such as breast, prostate, CRC and lung. In the future experiments should be performed under real time analysis.

REFERENCES

[1] Allard WJ, Matera J, Miller MC, Repollet M, Connelly MC, Rao C, et al. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. Clin Cancer Res. 2004;10:6897–904.
[2] Sun YF, Yang XR, Zhou J, Qiu SJ, Fan J, Xu Y. Circulating tumor cells: advances in detection methods, biological issues, and clinical relevance. J Cancer Res Clin Oncol. 2011;137:1151–73.
[3] Xenidis N, Ignatiadis M, Apostolaki S, Perraki M, Kalbakis K, Agelaki S, et al. Cytokeratin-19 mRNA-positive circulating tumor cells after adjuvant chemotherapy in patients with early breast cancer. J Clin Oncol. 2009;27:2177–84.
[4] Bidard FC, Mathiot C, Delaloge S, Brain E, Giachetti S, de Cremoux P, et al. Single circulating tumor cell detection and overall survival in nonmetastatic breast cancer. Ann Oncol. 2010;21:729–33.
[5] Cristofanilli M, Hayes DF, Budd GT, Ellis MJ, Stopeck A, Reuben JM, et al. Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. J Clin Oncol. 2005;23:1420–30.
[6] Liu MC, Shields PG, Warren RD, Cohen P, Wilkinson M, Ottaviano YL, et al. Circulating tumor cells: a useful predictor of treatment efficacy in metastatic breast cancer. J Clin Oncol. 2009;27:5153–9.

[7] Alunni-Fabbroni M. Circulating tumor cells in clinical practice: methods of detection and possible characterization. Methods. 2010;50:289–97.

[8] Bidard FC, Peeters DJ, Fehm T, Nole F, Gisbert-Criado R, Mavroudis D, et al. Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. Lancet Oncol. 2014;15:406–14.

[9] Gradilone A, Naso G, Raimondi C, Cortesi E, Gandini O, Vincenzi B, et al. Circulating tumor cells (CTCs) in metastatic breast cancer (MBC): prognosis, drug resistance and phenotypic characterization. Ann Oncol. 2011;22:86–92.

[10] Fehm T, Muller V, Aktas B, Janni W, Schneeweiss A, Stickeler E, et al. HER2 status of circulating tumor cells in patients with metastatic breast cancer: a prospective, multicenter trial. Breast Cancer Res Treat. 2010;124:403–12

[11] Hayes DF, Smerage J. Is there a role for circulating tumor cells in the management of breast cancer? Clin Cancer Res. 2008;14:3646–50.

[12] 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

[13] 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.

[14] A. Balmain, J. Gray, and B. Ponder, “The genetics and genomics of cancer,” Nature Genetics, vol. 33, no. 3, pp. 238–244, Mar. 2003.

[15] 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.

[16] 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.

[17] 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.

[18] G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou, “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes,” ClinicalCancer Res., vol. 14, no. 16, pp. 5158–5165, Aug. 2008.

[19] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: The next generation,” Cell, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[20] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe- Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. E. D. E. R. Schutz, D. R. Goldstein, M. Piccart, and M. Delorenzi, “Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures,” Breast Cancer Res., vol. 10, no. 4, p. R65, 2008

[21] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. A. Nuyten, A. B. Nobel, L. J. van’t Veer, and C. M. Perou, “Concordance among gene-expression based predictors for breast cancer,” N. Engl. J. Med., vol. 355, no. 6, pp. 560–569, Aug. 2006.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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

[27] Zhu and S. Kwong, “Gbest-guided artificial bee colony algorithm for numerical function optimization,” Applied Mathematics and Computation, vol. 217, no. 7, pp. 3166–3173, 2010.

[28] W.-F. Gao and S.-Y. Liu, “A modified artificial bee colony algorithm,” Computers and Operations Research, vol. 39, no. 3, pp. 687–697, 2012.

[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.