ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2018.0244

DEEP LEARNING FEATURE EXTRACTION WITH ENSEMBLE SPECTRAL CLUSTER AND GAUSSIAN MIXTURE FOR MALICIOUS TUMOR DETECTION

S. Subash Chandra Bose[1] and T. Christopher[2]
[1]*Department of Computer Science, Government Arts College, Udumalpet, India*
[2]*Department of Computer Science, Government Arts College, Coimbatore, India*

Abstract

Different clustering algorithms produce distinct sub-divisions as they apply disparate partition on the data. Hence, no single clustering algorithm is said to be optimal and therefore resulting in different partitions. To utilize the complementary nature of different partitions, ensemble clustering is used. The work in this paper focuses on producing ensembles through several clustering algorithms that perform feature extraction using deep learning and malicious tumor detection through ensemble cluster. In this study, to improve the performance and reduce the complexity involved in the malicious tumor detection process, Deep Learning Feature Extraction (DLFE) technique is presented. Furthermore, to improve the quality of results obtained, ensemble clusters namely, Normalized Spectral Cluster and Gaussian Mixture technique has been applied to the extracted features. The experimental results of the proposed technique have been evaluated and validated for performance and quality analysis on three datasets based on accuracy, sensitivity, specificity. The experimental results achieved 85.28% accuracy, 70.43% specificity, and 97.19% sensitivity, demonstrating the effectiveness of the proposed technique for identifying normal and abnormal tissues from various test images. The simulation results prove the significance in terms of quality parameters and accuracy in comparison to the state-of-the-art techniques.

Keywords:

Clustering Algorithm, Deep Learning, Feature Extraction, Normalized Spectral Cluster, Gaussian Mixture

1. INTRODUCTION

Recently, the ceaseless development of microarray methods and their applications in the field and research of cancer provides a new avenue towards treatment and diagnosis of cancer. There has been an increasing interest in providing measures for discovering the underlying classes from cancer gene expression because of its important applications in cancer diagnosis, treatment and related areas. Tumor clustering plays a pivotal role in obtaining the malignancies from cancer gene expression.

Random Forest using Class Decomposition (RF-CD) method was investigated in [1] for medical diagnosis. The method using Random Forest was found to be applied in any classification method, including single classifier system. To start with, k-means clustering was applied to instances that belong to each class by varying number of clusters.

Once each class was disintegrated in its subclasses, a Random Forest was applied to the newly class-engineered data set. This process was said to be performed in an iterative manner. The method exhibited higher amount of accuracy. In spite of higher amount of accuracy being observed, the dimension factor was less concentrated.

Feature Selection-based Semi Supervised Cluster Ensemble (FS-SSCE) framework for clustering of tumor cells from bio molecular data was investigated in [2] that featured two properties. First, FS-SSCE framework adopted feature selection techniques to eliminate the effect of noisy genes. Second, the deployment of binate constraint based K-means algorithm considered the experts' knowledge effect. Finally, a double selection based semi-supervised cluster ensemble framework (DS-SSCE) was formed.

The DS-SSCE framework applied feature selection technique to perform gene selection on gene dimension. It also selected an optimal subset of representative clustering solutions with the objective of improving the tumor clustering performance using normalized cut algorithm. An optimal subset of clustering solutions was obtained by adopting multiple clustering solution selection strategies. Despite, improvement obtained in tumor clustering performance, quality of results remained unaddressed. Unfortunately, the clustering step is laborious owing to the dimensionality factor and quality of results obtained through clustering. Moreover, the existing clustering methods focused on clustering accuracy by performing feature selection methods. Therefore, the primary aim of this paper is to develop an improved ensemble clustering technique for effective malicious tumor detection by concentrating on the quality of clusters produced.

The rest of the paper is organized as follows: section 2 presents the related works. Section 3 presents the methods with the steps used in the proposed technique. Section 4 presents the experimental settings with performance metrics. Section 5 presents the comparative analysis and detailed discussion with the aid of table and graph. Finally, section 6 contains the conclusions.

2. RELATED WORKS

In an attempt to attain enhanced classifier accuracy, substantial research has been administered in classifier ensembles. Cluster ensembles have only unfolded very recently. Clustering ensembles can be generated in different ways, like producing ensembles through several clustering algorithms, running same algorithm with different parameters or by using different samples.

An integrated robust semi-supervised framework using ensemble method for heterogeneous datasets was investigated in [3]. The focus of the framework remained in improving the predictive capacity. Yet another clustering-based ensemble method based on weighted One Class Support Vector Machines (OCSVM) [4] resulted in the improvement of computational complexity.

A consensus based cluster ensemble framework integrating fuzzy extension model was presented in [5] resulting in the

improvement of tumor clustered data. Machine learning methods were investigated in [6] with the objective of reducing the human workload. The analysis of gene expression data has paramount applications for treatment related to gene, cancer diagnosis and so on. In [7], a method called Projective Clustering Ensemble was presented to improve the clustering quality of gene expression data. This was said to be achieved by combining multiple projective clustering.

In several areas it has been shown that a cluster ensemble is frequently found to be more precise than any of the single clustering techniques. This has led to the further investigation in ensemble techniques for clustering. An ensemble clustering algorithm for clustering cancer data using hierarchical clustering was presented in [8] [9], ensuring accuracy of data produced. However, relevancy was not ensured. To address this issue, an agent-based algorithm was presented in [10] ensuring consistency and interpretable collection of clusters.

Despite humans being excellent cluster seekers in two and three dimensions, however, automatic algorithms are required for high-dimensional data. In [11], two-density k-means algorithm was investigated to address issues related to high-dimensional data. A spectral clustering algorithm in social networks was designed in [12].

Two link prediction methods, based on k-medoids and landmark was presented resulting in higher accuracy. Ensemble cluster in the field of tumor detection is receiving greater attention. In [13], pre-segmented and post-segmented method was presented to extract the desired feature. Followed by it, a hybrid feature block was presented to show effective computer-aided diagnosis performance.

In recent years, with the advent of information technology and e-health care system in the medical field, experts in the field of clinical arena are provided with better health care to the patient. To enhance the performance and minimize the complexity involved in brain tumor detection, Berkeley Wavelet Transform and Support Vector Machine were investigated in [14]. This resulted in the improvement of accuracy and cluster quality. However, a unified cluster structure from multiple cluster structure from different datasets remained unaddressed. To address this issue, Distribution-based Cluster Structure Ensemble (DCSE) framework was presented in [15] to further improve the cluster quality and therefore the rate of tumor detection.

Multi-class clustering for gene marker identification through SVM-based ensemble was investigated in [16] for detecting multiple cancer subtypes. Yet another ensemble of classifiers with the generated cluster was presented in [17] for detecting remotely located data points to cope with newer situation. However, continuous optimization with the same local optima remained an unsolved issue. To this, a randomized greedy modularity algorithm was presented in [18] that not only found local optimal solutions, but also ensured accuracy and tumor detection rate.

Due to the large number of possibilities in gene expression data, selecting the most effective clustering method, for a specific set of gene expression data, is the highly preferred one. Despite, several research works conducted using hierarchical clustering, it appears to be sub-optimal. To improve the robustness and quality of clustering result, link-based cluster ensemble method was presented in [19].

The above literature survey has disclosed that some of the methods were invented to acquire segmentation only, whereas some of the methods were invented to acquire feature extraction and some of the methods were presented to acquire clustering only. In addition, only few features were extracted and therefore lesser accuracy in tumor detection has been evolved. In this study, we perform a combination of deep learning as a feature extraction method and spectral cluster with Gaussian mixture as an ensemble cluster method to deeply extract features and improve cluster quality. The cause of this study is to extract relevant features from the testing dataset and cluster malicious and non-malicious tumor regions.

3. DEEP LEARNING FEATURE EXTRACTION WITH SPECTRAL CLUSTER AND GAUSSIAN MIXTURE

The major steps of the proposed technique, Deep Learning feature extraction with Spectral Cluster and Gaussian Mixture (DL-SCGM) comprise of extraction of invariant features and perform ensemble clusters for malicious tumor detection. The two-step process technique is graphically represented by Fig.1 and the two phases are explained in the subsequent sections.

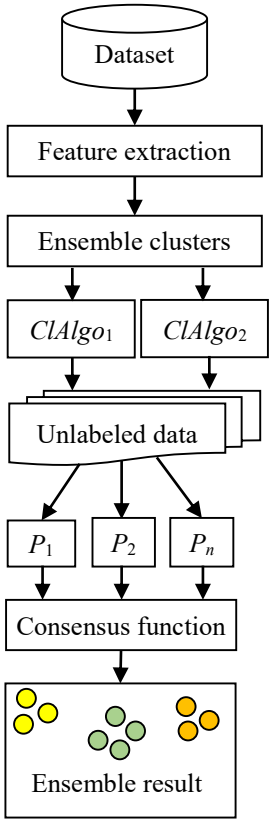

Fig.1. Block diagram of DL-SCGM

As shown in the above Fig.1, ensemble clusters of spectral cluster and Gaussian mixture is used for malicious tumor detection. To start with a dataset provided as input, relevant and reduced features are extracted using Deep Feature Synthesis. Followed by the extracted features, an ensemble of cluster namely, Normalized Spectral Cluster and Gaussian Mixture for malicious tumor detection is presented.

As illustrated in the above figure, with the reduced features extracted using Deep Feature Synthesis, features an input dataset cluster algorithm 1 represented by $ClAlgo_1$ denotes the normalized spectral cluster and cluster algorithm 2 represented by $ClAlgo_2$ denotes the Gaussian mixture for malicious tumor detection, constitutes ensemble clusters. With the aid of ensemble clusters, the proposed work combines multiple partitions of extracted features into a single partition to ensure quality of results achieved.

Finally, ensemble results are obtained with the aid of a consensus function based on probability likelihood function.

3.1 DEEP LEARNING FEATURE EXTRACTION TECHNIQUE

This paper deals with the most trivial issues of relevant feature extraction for malicious tumor detection. Often, methods involved in feature extraction believe in certain robust criterion in search of lower dimensional representation. However, the true structure of the data is said to be unknown. This in turn makes it laborious to define a suitable criterion. A new feature extraction method is presented in this work incorporating the idea of deep learning.

To test the performance of the proposed technique, Leukemia, Lung Cancer and Breast Tissue dataset are considered. The input of DLFE technique comprises of Leukemia, Lung Cancer and Breast Tissue dataset and the output being the feature extracted through deep learning. In Fig.as an example, the first layer, there are two feature maps whereas in the second layer, there include four feature maps.

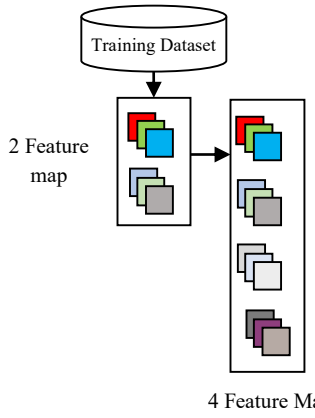

Fig.2. Block diagram of DLFE

Deep Learning Feature Extraction (DLFE) technique involves the process of collecting higher level information involving several instances and attributes such as impedance measurements, characteristics of cell nuclei present in the image and so on. Feature analysis is an important parameter of human visual perception that is used to enhance the accuracy of diagnosis and tumor detection system by selecting prominent features. The DLFE technique follows two steps for feature extraction from the given input training dataset.

In the first step, the modeling of features using greedy technique is formulated. In the second step, to obtain the relevant features and eliminate redundant features, mean activation function is employed. The pseudo code representation of Greedy Deep Activated feature extraction algorithm is given below.

Algorithm 1: Greedy Deep Activated feature extraction algorithm

Input: Features '$F=f_1,f_2,\ldots,f_n$'

Output: Features extracted

1: Begin
2: For each Features F
3 Obtain deep hierarchical representation using Eq.(1)
4 Measure activation function using Eq.(2)
5: Return extracted features (EF)
6: End for
7: End

This Greedy Deep Activated feature extraction algorithm is trained using greedy method by a deep hierarchical representation of the training dataset. The mathematical representation to model features for n hidden layers HL^n is as given below.

$$P\left(F, HL^n\right) = \sum_{i=1}^{n} P\left(HL^i \middle| HL^{i+1}\right) * P\left(HL^{n-2} \middle| HL^n\right) \quad (1)$$

The training of DLFE uses a greedy layer-wise method on the extracted invariant features. The first layer consists of input of features. With this layer, the second layer is extracted by providing the training examples as $P(HL^1 \mid HL^0)$. Next, the second layer is trained using mean activation function of training dataset. This step is iterated for desired number of layers upward the mean values of the activation function. Let us consider a pre-defined score value with '*score*'. Then, the activation function for each extracted feature is as given below.

$$AF\left(EF\right) = \begin{cases} 0 & for\ F < score \\ 1 & for\ F > score \end{cases} \quad (2)$$

From Eq.(2), the correct feature classes extracted possess higher score than the other feature classes. Hence, the mean values of activation function remain the concatenated features with greater score values.

3.2 ENSEMBLE CLUSTERS USING NORMALIZED SPECTRAL AND GAUSSIAN MIXTURE

With the resultant features extracted using the DLFE technique, the next step is to perform an ensemble cluster for malicious tumor detection. The issue of the quality of results obtained and its corresponding tumor detection accuracy poses severe problems. To this, an ensemble cluster using Normalized Spectral and Gaussian Mixture is presented in this work.

Normalized Spectral Clustering is applied to the extracted features with the objective of obtaining clusters of higher quality. With the clusters of higher quality being obtained, the resultant of spectral cluster includes both malicious and non-malicious tumor region, also called as mixture region. Therefore, to ensure a better rate of malicious tumor detection, a Gaussian Mixture technique is applied to the resultant spectral cluster.

The Fig.3 shows the flow diagram of the Normalized Spectral Clustering technique. As shown in the figure, the proposed Normalized Spectral Cluster technique applies a clustering technique to the extracted features based on the new similarity measure, called Twin Similarity Matrix. The Twin Similarity

Matrix is designed in such a way to improve the quality of results and therefore the detection rate through Gaussian Mixture technique.

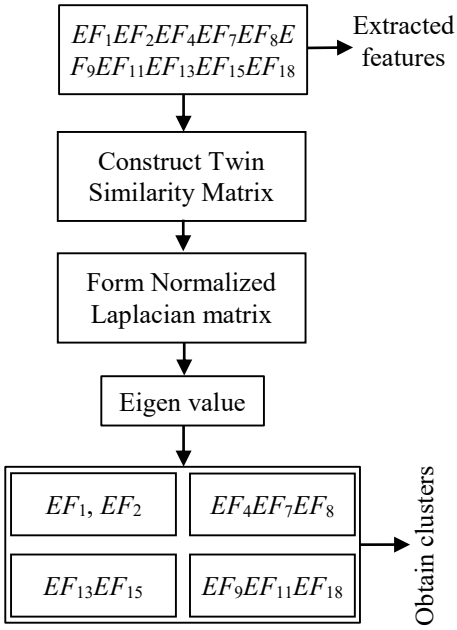

Fig.3. Flow diagram of Normalized Spectral Cluster

Normalized Laplacian Matrix schema is used to find an optimal transformation of features in each obtained cluster to measure the respective centers. The respective centers obtained finally denote the clusters used to characterize the patterns and therefore detect the malicious tumor. The detection of a malicious tumor is performed via Gaussian Mixture technique. The pseudo code representation of Twin Normalized Clustering algorithm is given below.

Algorithm 2: Twin Normalized Clustering algorithm

Input: Extracted Features $EF = ef_1, ef_2,\ldots, ef_n$, Clusters '$m$', Threshold '$T$'

Output: Cluster

1: Begin
2: For each Extracted Features 'EF'
3: Measure strength of clustering using Eq.(3)
4: If $ST_a^p > T$ then
5: Select corresponding cluster CL_a^p
6: Else
7: Go to Step 3
8: End if
9: Form Twin Similarity Matrix using Eq.(4)
10: Form Symmetric Laplacian Matrix using Eq.(5)
11: Obtain Eigen Values using Eq.(6)
12: Obtain Cluster Matrix using Eq.(7) and obtain resultant clusters
13: End for
14: End

Let us consider n features extracted using DLFE technique. The first step in Normalized Spectral Clustering technique is to form Twin Similarity Matrix. With the objective of improving the quality of clustered results, the quality of each cluster CL_i^j is assessed based on the cluster strength, resulting in tuples '$\left(CL_i^j, ST_i^j\right)$', where i=1,2,…,m and j=1,2,…,n. Here m represents the number of clusters in the partition produced by Greedy Deep Activated feature extraction algorithm.

Relevant clusters are filtered out from these partitions, by choosing only those clusters that manifest mean strength greater than a given threshold, T. The strength of individual clusters within each partition is evaluated as the mean strength of the twins of features in each cluster. It is mathematically evaluated as given below,

$$ST_a^p = \frac{ST_a^p(i,j)}{\left|CL_a^p\right|\left(CL_a^p - 1\right)}. \quad (3)$$

From Eq.(3), the strength ST for a^{th} cluster, with p partition, is obtained using the number of features in cluster CL_a^p. The Twin Similarity Matrix is then mathematically formulated as given below,

$$TSM_{ij} = \exp\left(Dis\left(CL_i, CL_j\right)\right). \quad (4)$$

From above the twin similarity matrix TSM_{ij} is measured on the basis of the exponentiation exp of the distance Dis between the two closer or similar clusters CL_i and CL_j respectively. With the obtained Twin Similarity Matrix, a Normalized Laplacian Matrix is formulated as given below,

$$NLM_{ij} = D_{ij} - TSM_{ij}. \quad (5)$$

From Eq.(5), D_{ij} represents the degree matrix for i^{th} row and j^{th} column. Followed by this, the Eigen values EV is obtained for the Normalized Laplacian Matrix with the 2×2 identity matrix I_2 as given below,

$$EV_{ij} = (\lambda I_2 - NLM_{ij}). \quad (6)$$

Let EV_{ij} represents the Eigen vector for i^{th} row and j^{th} column. Finally, the Normalized Spectral Clustering technique clusters the features from Eigen vector EV_{ij} with selection of significant clusters based on the strength of the cluster using maximal rule.

$$CM_{ij} = \max\left\{EV_{ij}\left[CL_a^p, ST_a^p\right]\right\} \quad (7)$$

From Eq.(7), a maximal rule combines the eigen vector and forms a cluster matrix CM_{ij} with the aid of the a^{th} cluster with p^{th} partition having higher strength ST than that of the threshold respectively. With this, more similar features are said to form a cluster. However, the cluster may include both malicious and/or non-malicious tumor cells. With the objective of detecting the malicious tumor, the proposed work, applies Gaussian Mixture technique for early detection of tumor that is discussed in the forthcoming section.

3.3 GAUSSIAN MIXTURE MALICIOUS TUMOR DETECTION

The clusters obtained using Normalized Spectral Clustering technique as discussed in the above section includes both tumor and non-tumor cells. For early detection of malicious tumor, a Gaussian Mixture technique is applied. This GM technique characterizes between tumor and non-malicious tumor through probability distance model. The pseudo code representation of Gaussian Mixture Malicious Tumor Detection is as given below.

Algorithm 3: Gaussian Mixture Malicious Tumor Detection algorithm

Input: Cluster Matrix CM_{ij}

Output: Malicious tumor detection

1: Begin
2: For each Cluster Matrix CM_{ij}
3: Obtain the Gaussian Matrix using Eq.(8)
4: Obtain the Gaussian Matrix for 'd' dimension using Eq.(9) and Eq.(10)
5: Measure probability likelihood function using Eq.(12)
6: Obtain malicious class M_2 and non-malicious class M_1 using Eq.(14) and Eq.(15)
7: End for
8: End

Let us denote the parameters of Gaussian element as $\varphi_{ij} = \{\phi_{ij}, \mu_{ij}, CM_{ij}\}$, where ϕ_{ij}, μ_{ij} and CM_{ij} represents the mixing coefficient value, mean vector and cluster matrix respectively. Then, the entire Gaussian Mixture is written as given below,

$$\psi = \{n, \varphi_{ij}, \ldots, \varphi_{in}\}. \quad (8)$$

From Eq.(8), n represents the number of clusters in cluster matrix CM_{ij} and with this, the entire Gaussian Mixture technique for d dimension cluster CL is written as given below,

$$P(CL,\psi) = \sum_{i,j=1}^{m,n} \phi_{ij}\, p\left(cl, \mu_{ij}, CM_{ij}\right). \quad (9)$$

$$P(CL,\psi) = \phi_{ij} \frac{\exp\left(\left(cl - \mu_{ij}\right) * \left(CM_{ij}^{-1} - \mu_{ij}\right)\right)}{2\pi^{d/2}} \quad (10)$$

Finally, to characterize an edge between normal and malicious tumors, a probability likelihood function is used in the proposed work for early detection of malicious tumors. Let us define a likelihood function as $Prob(\{CL_F\}|\{M_F\})$, for the probability of observed cluster CL_{EF} conditioned on model functionalities M_{EF} of extracted features EF. With the following assumption,

$$p = (M_2|CL_1, CL_2\text{-}CL_1, M_1) \quad (11)$$

The probability likelihood function is then deduced as given below,

$$= p(M_2|CL_2\text{-}CL_1, M_1) \quad (12)$$

Then, the presence of malicious and non-malicious classes from the normalized spectral clustered function is then given as below,

$$Prob(M_1, M_2 \mid CL_1, CL_2). \quad (13)$$

where, M_1 = non-malicious class, M_2 = malicious class

$$= \sum_{i,j=1}^{n} \left(M_1 = b_i, M_2 = b_j \middle| CL_1, CL_2\right) \quad (14)$$

$$\sum_{i,j=1}^{n} p\left(M_2 = b_j \middle| CL_1, CL_2 - CL_1, M_1 = b_i\right), \left(M_1 = b_i \middle| CL_1\right) \quad (15)$$

The malicious and non-malicious cluster identified with the Gaussian Mixture technique makes the ensemble cluster the default choice for clustering of malicious and non-malicious tumor. With the ensemble cluster, malicious tumor detection is observed at an early stage. The Ensemble Cluster algorithm's performance is evaluated in terms of clustering time, clustering accuracy, sensitivity, and specificity.

4. PERFORMANCE EVALUATION (JAVA)

To validate the performance of our DL-SCGM technique, three benchmark datasets were used called, Breast tissues (BT), WDBC (Wisconsin Diagnostic Breast Cancer) data sets Lung cancer (LC) [20] and Leukemia [21] respectively, which included sample images of 100 patients. The first dataset is the Breast tissues (BT).

For the purpose of analysis, we considered 100 instances which included both malicious and non-malicious tumor classes. The second dataset is the WDBC (Wisconsin Diagnostic Breast Cancer) data sets Lung cancer (LC) which consists of 569 number of instances of which experiments is performed with 100 different instances. Finally, the third dataset is the Leukemia that is also used to compare the results of our proposal with the state-of-the-art works.

The leukemia dataset was taken from a collection of leukemia patient samples reported by Golub. The leukemia dataset serves as benchmark for microarray analysis methods. This dataset comprises of measurements corresponding to acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML) samples from bone marrow and peripheral blood.

The dataset consisted of 72 samples: 25 samples of AML, and 47 samples of ALL. Each sample is measured over 7,129 genes. On the other hand, the Lung Cancer dataset comprises of classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA).

The training set contains 32 of them, 16 MPM and 16 ADCA. The rest 149 samples are used for testing. Each sample is described by 12533 genes. The validation of the DL-SCGM technique was done using the metrics, tumor clustering accuracy, tumor clustering time, sensitivity and specificity. Performance metrics are expressed by the following equations.

The tumor clustering accuracy rate measures the ratio of number of malicious and non-malicious tumor correctly clustered to the total number of sample cases used during experimentation. This is mathematically formulated as given below,

$$Accuracy_{TC} = \frac{No.of\ M_1\ and\ M_2\ correctly\ clustered}{Total\ number\ of\ sample\ cases}. \quad (16)$$

From Eq.(16), $Accuracy_{TC}$ refers to the tumor clustering accuracy with M_1 and M_2 denoting the non-malicious and malicious class respectively. Higher clustering accuracy indicates the effectiveness of the method. The tumor clustering time is defined as the time taken to cluster the tumor with respect to the total number of sample cases used during experimentation. This is mathematically formulated as given below,

$$Time_{TC} = Time(Clustering[M_1] * Clustering[M_2]) * n. \quad (17)$$

From Eq.(17), $Time_{TC}$ refers to the tumor clustering time with M_1 and M_2 denoting the non-malicious and malicious class respectively. Lower clustering time indicates the effectiveness of the method. Sensitivity also referred to as the positive rate, measures the ratio of malicious class correctly identified as having the condition. Specificity also referred to as the true negative rate, measures the ratio non-malicious that are correctly identified as such.

5. DISCUSSION

This section presents the results of our proposed ensemble cluster technique, which are obtained using Leukemia, Lung Cancer and Breast Tissue dataset. The proposed algorithm was carried out using JAVA, which runs on the Windows 8 operating system and has an Intel core i3 processor and a 4GB RAM.

5.1 COMPARATIVE ANALYSIS AND DISCUSSIONS BASED ON LEUKEMIA DATASET

The Table.1 given below provides the details of the different performance metrics such as tumor clustering accuracy, tumor clustering time, sensitivity and specificity.

Table.1. Performance metrics based on Leukemia dataset

Metrics	DL-SCGM	RF-CD	DS-SSCE
Tumor Clustering accuracy (%)	76.32	68.13	61.23
Tumor Clustering time (ms)	66.32	74.33	82.34
Sensitivity (%)	72.14	68.43	64.12
Specificity (%)	68.13	62.83	58.90

A higher value of tumor clustering accuracy and lower value of tumor clustering time indicate better quality of results in the extracted features. These metrics represented the highest performance using DL-SCGM technique with the clustering accuracy observed to be 76.32% and clustering time being 66.32ms while experimentation performed using 50 samples.

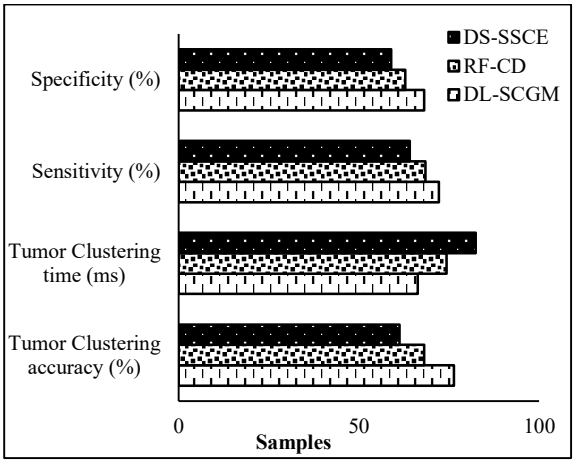

Fig.4. Comparative analysis of DL-SCGM, RF-CD and DS-SSCE using Leukemia dataset

The Fig.4 as given above shows the performance of DL-SCGM, RF-CD and DS-SSCE techniques with respect to tumor clustering accuracy, tumor clustering time, sensitivity and specificity. As mentioned earlier, in order to ensure consistency of the results, the experiment was repeated 10 times for each set, and the results were averaged and 50 samples were used for experimentation. Here, we also compare the DL-SCGM technique to RF-CD and DS-SSCE. It is observed that DL-SCGM technique significantly outperforms the state-of-the-art works. For example, DL-SCGM technique obtains the best average tumor clustering accuracy of 76.32% and tumor clustering time of 66.32ms on Leukemia dataset. The possible reasons are as follows. Compared with other cluster ensemble approaches, the experts' knowledge in the form of deep learning feature extraction is now considered, which provides extraction of invariant features to facilitate clustering. Next, compared with RF-CD that applies class decomposition, DL-SCGM technique integrates multiple clustering solutions with the aid of Normalized Spectral Clustering and Gaussian Mixture to improve the performance of class decomposition and obtains more accurate results. Finally, compared with DS-SSCE, Greedy Deep Activated feature extraction algorithm is adopted for selecting prominent features to eliminate the unwanted features.

5.2 COMPARATIVE ANALYSIS AND DISCUSSIONS BASED ON LUNG CANCER DATASET

The test performance of the ensemble clusters using Normalized Spectral Cluster and Gaussian Mixture determined by the computation of the statistical parameters such as tumor clustering accuracy, tumor clustering time, sensitivity and specificity, based on Lung Cancer dataset with different clustering techniques is shown in Table.2.

Table.2. Performance metrics based on Lung Cancer dataset

Metrics	DL-SCGM	RF-CD	DS-SSCE
Tumor Clustering accuracy (%)	85.28	81.19	79.26
Tumor Clustering time (ms)	70.62	74.86	78.29
Sensitivity (%)	97.19	94.32	92.84
Specificity (%)	44.94	42.58	39.72

As shown in the table, higher values of accuracy and sensitivity and a lower value of specificity indicate better performance. It can be seen from Table.2 that the performance of our Twin Normalized Clustering algorithm is better than the state-of-the-art techniques.

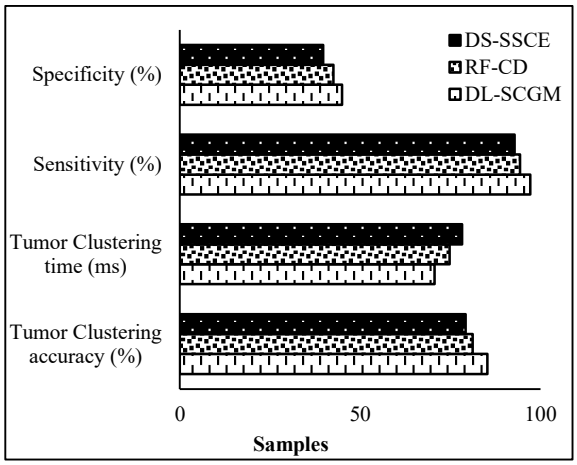

Fig.5. Comparative analysis of DL-SCGM, RF-CD and DS-SSCE using Lung Cancer dataset

The Fig.5 shows the performance of DL-SCGM, RF-CD and DS-SSCE using Lung Cancer dataset. As shown in the figure, a comparative study was employed using ensemble clusters. Ensembles namely, class decomposition and semi-supervised clustering ensemble method were considered [1] [2]. Random

forest and Double selection were evaluated. Tumor clustering accuracy based on the DL-SCGM technique showed the highest accuracy value of 85.28%, 81.19% by applying RF-CD and 79.26% by applying DS-SSCE respectively using Lung Cancer dataset. From the figure, the tumor detection rate was improved using Lung Cancer dataset with the best performance achieved using DL-SCGM technique, where the sensitivity was found to be 97.19% and specificity was observed to be 44.94%. In DL-SCGM technique, ensemble clusters showed highest value (85.28%, 70.62%, 97.19% and 44.94%) of tumor clustering accuracy, tumor clustering time, sensitivity and specificity respectively.

We note that the main goal of the feature extraction using Deep Learning technique is that the greedy technique and mean activation function usually contain the most important features (lower dimensional representation).

Hence they constitute one part of the extracted features and another represented by the critical features from the twin similarity matrix. In comparison, DL-SCGM technique is more promising using Normalized Laplacian matrix formed via twin similarity matrix with the highest tumor clustering accuracy and lower tumor clustering time.

5.3 COMPARATIVE ANALYSIS AND DISCUSSIONS BASED ON BREAST TISSUE DATASET

The performance of the proposed DL-SCGM technique is also compared with two state-of-the-art ensemble clustering techniques, RF-CD [1] and DS-SSCE [2] by using Breast Tissue dataset. The obtained results were reported in Table.3.

Table.3. Performance metrics based on Breast Tissue dataset

Metrics	DL-SCGM	RF-CD	DS-SSCE
Tumor Clustering accuracy (%)	79.27	75.35	72.58
Tumor Clustering time (ms)	63.02	67.03	76.17
Sensitivity (%)	76.28	72.38	70.30
Specificity (%)	70.43	57.53	54.68

In contrast with RF-CD [1] and DS-SSCE [2], the features selecting processing is automatic but the training processing involves a time consuming task, as it does not consider cluster proximity between different clusters. Therefore, the DL-SCGM technique has combined the processing of feature extraction with deep learning that considers mean activation function in extracting the relevant features. The proposed technique took comprehensive lead of deep features from the Greedy Deep Activated feature extraction algorithm, resulting in a synthesized feature extraction structure.

The Fig.6 shows the comparative analysis of DL-SCGM, RF-CD and DS-SSCE using Breast Tissue dataset. In the case of the DS-SSCE as presented in [2], the statistical values are obtained such as tumor clustering accuracy of 72.58%, tumor clustering time of 76.17ms, sensitivity of 70.30% and specificity of 54.68%. However, the ensemble clustering algorithm as used in RF-CD [1] provided better performance (75.35%, 67.03%, 72.38% and 57.53%) compared to DS-SSCE. The significance results of the DL-SCGM technique is also displayed as shown in Fig.6. As shown in this figure, the characteristic curve representing four performance metrics of the proposed DL-SCGM technique has significantly advanced the performance of the differentiation between malicious and non-malicious tumor compared to the other two ensemble clusters. The DL-SCGM technique has performed well due to the fact that it selected the most prominent features. With these prominent features relevant clusters were filtered by selecting those clusters that had a mean strength value greater than the threshold value. Followed by this, the significant cluster selection was made using the maximal rule.

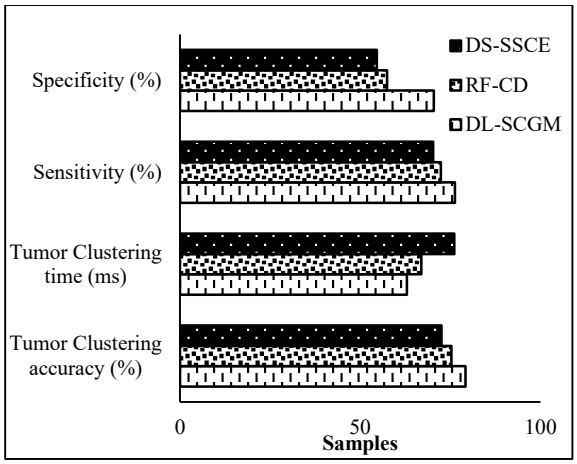

Fig.6. Comparative analysis of DL-SCGM, RF-CD and DS-SSCE using Breast Tissue dataset

6. CONCLUSION

In this work, a Deep Learning feature extraction technique with Ensemble Clusters for malicious tumor detection is analyzed. To extract relevant and invariant features deep learning technique were used. With the relevant extracted features, a normalized spectral clustering technique was applied using a twin matrix clustering algorithm based on a threshold technique to improve cluster quality and therefore characterize the patterns for efficient malicious tumor detection. Furthermore, we used Gaussian Mixture Malicious Tumor Detection algorithm to obtain similar mixture pattern and probability likelihood function to detection malicious tumor by analyzing observed cluster and model functionalities. From the experimental results performed on different datasets, it is clear that the analysis for malicious tumor detection is accurate when compared with the state-of-the-art works. Various performance factors also indicate that the proposed algorithm provides a better result by improving certain parameters such as tumor clustering accuracy, time, sensitivity and specificity. Our experimental results show that the proposed technique can aid in the timely detection of a malicious tumor.

REFERENCES

[1] Eyad Elyan and Mohamed Medhat Gaber, “A fine-Grained Random Forests using Class Decomposition: An Application to Medical Diagnosis”, *Neural Computing and Applications*, Vol. 27, No. 8, pp. 2279-2288, 2015.

[2] Zhiwen Yu, Hongsheng Chen Jane You, Hau-San Wong, Jiming Liu, Le Li and Guoqiang Han, “Double Selection based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles”, *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics, Vol. 11, No. 4, pp. 1113-1119, 2014.

[3] Smita Prava Mishra, Debahuti Mishra and Srikanta Patnaik, "An Integrated Robust Semi-Supervised Framework for Improving Cluster Reliability using Ensemble Method for Heterogeneous Datasets", *Karbala International Journal of Modern Science*, Vol. 1, No. 4, pp. 200-211, 2015.

[4] Bartosz Krawczyk, Michal Wozniak and Boguslaw Cyganek, "Clustering-based Ensembles for One-Class Classification", *Information Sciences*, Vol. 264, pp. 182-195, 2014.

[5] Zhiwen Yu, Hantao Chen, Jane You Jiming Liu, Hau-San Wong and Guoqiang Han, "Adaptive Fuzzy Consensus Clustering Framework for Clustering Analysis of Cancer Data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 12, No. 4, pp. 1123-1129, 2015.

[6] Yannik Siegert, Xiaoyi Jiang, Volker Krieg and Sebastian Bartholomaus, "Classification-Based Record Linkage with Pseudonymized Data for Epidemiological Cancer Registries", *IEEE Transactions on Multimedia*, Vol. 18, No. 10, pp. 224-237, 2016.

[7] Xianxue Yu, Guoxian Yu and Jun Wang, "Clustering Cancer Gene Expression Data by Projective Clustering Ensemble", *PLOS ONE*, Vol. 12, No. 2, pp. 1-21, 2017.

[8] Ran Qi, Dengyuan Wu, Li Sheng, Donald Henson, Arnold Schwartz, Eric Xu, Kai Xing and Dechang Chen, "On an Ensemble Algorithm for Clustering Cancer Patient Data", *BMC System Biology*, Vol. 7, No. 4, pp. 1-9, 2013.

[9] Filippo Maria Bianchi, Lorenzo Livi and Cesare Alippi, "Investigating Echo-State Networks Dynamics by Means of Recurrence Analysis", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 2, pp. 81-85, 2016.

[10] Filippo Maria Bianchi, Enrico Maiorino, Lorenzo Livi, Antonello Rizzi and Alireza Sadeghian, "An Agent-based Algorithm Exploiting Multiple Local Dissimilarities for Clusters Mining and Knowledge Discovery", *Soft Computing*, Vol. 21, No. 5, pp. 1347-1369, 2015.

[11] Filippo Maria Bianchi, Lorenzo Livi and Antonello Rizzi, "Two Density-based k-means Initialization Algorithms for Non-Metric Data Clustering", *Pattern Analysis and Applications*, Vol. 19, No. 3, pp. 745-763, 2015.

[12] Asmaa M. Mahmoud, Lamiaa M.E. Bakrawy and Neveen I. Ghali, "Link Prediction in Social Networks based on Spectral Clustering using k-Medoids and Landmark", *International Journal of Computer Applications*, Vol. 168, No. 7, pp. 1-8, 2017.

[13] Bushra Mughal, Muhammad Sharif and Nazeer Muhammad, "Bi-Model Processing for Early Detection of Breast Tumor in CAD System", *The European Physical Journal Plus*, Vol. 16, No. 4, pp. 132-136, 2017.

[14] Nilesh Bhaskarrao Bahadure, Arun Kumar Ray and Har Pal Thethi, "Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction using Biologically Inspired BWT and SVM", *International Journal of Biomedical Imaging*, Vol. 2017, pp. 1-12, 2017.

[15] Zhiwen Yu, Xianjun Zhu, Hau-San Wong, Jane You, Jun Zhang and Guoqiang Han, "Distribution-Based Cluster Structure Selection", *IEEE Transactions on Cybernetics*, Vol. 47, No. 11, pp. 3554-3567, 2016.

[16] Anirban Mukhopadhyay, Sanghamitra Bandyopadhyay and Ujjwal Maulik, "Multi-Class Clustering of Cancer Subtypes through SVM Based Ensemble of Pareto-Optimal Solutions for Gene Marker Identification", *PLOS ONE*, Vol. 5, No. 11, pp. 1-7, 2010.

[17] Mohammad Raihanul Islam, Md. Mustafizur Rahman, Asif Salekin and Ahmed Shayer Andalib, "A Novel Approach for Generating Clustered Based Ensemble of Classifiers", *International Journal of Machine Learning and Computing*, Vol. 3, No. 1, pp. 137-141, 2013.

[18] Andreas Geyer-Schulz and Michael Ovelgonne, "*The Randomized Greedy Modularity Clustering Algorithm and the Core Groups Graph Clustering Scheme*", German-Japanese Interchange of Data Analysis Results, 2014.

[19] Natthakan Iam-On, Tossapon Boongoen and Simon Garrett, "LCE: A Link-Based Cluster Ensemble method for Improved Gene Expression Data Analysis", *Bioinformatics*, Vol. 26, No. 12, pp. 1513-1519, 2010.

[20] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation", *Proceedings of the National Academy of Sciences*, Vol. 96, No. 6, pp. 2907-2912, 1999.

[21] A. Bhattacharjee, W.G. Richards and J. Staunton, "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinomas Sub-Classes", *Proceedings of the National Academy of Sciences*, Vol. 98, No. 24, pp. 13790-13795, 2001.