ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0386

MRI-BRAIN TUMOR CLASSIFICATION USING K-MEANS CLUSTERING AND ENHANCED HARMONY FEATURE SELECTION

B. Sathees Kumar
Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, India

Abstract

This study introduces an enhanced feature selection method that is efficient in differentiating the malignant tumor patients from the benign patients by using K-Means clustering method combined with enhanced harmony search algorithm. The start of malignant tumor is caused by gene mutation process, it is very vital to identify and classify the presence or absence of the malignant tumor through analyzing the gene information. The planned methodology composed of four steps. The first step is to preprocess the original data by using min-max normalization. In the next step, generalized fisher score is used to find and eliminate the redundant data to confine the significant candidate genes. Selection of representative gene from each cluster is done by the K-Means clustering technique in the next phase. In the final phase the vital features for classification are selected by enhanced harmony search algorithm. The selected gene combination through this method for feature selection is then applied to the classification model and verified by means of 5-fold cross validation method. The projected model obtained a classification accuracy of up to 96.67%. Additionally, on comparing the projected method with other methods, the projected method performs well in classifying malignant tumor. This new method performs well in classification of brain tumors to malignant or benign. The projected model cannot be restricted only with the classification of brain tumors, but can also be used for other gene-related diseases effectively.

Keywords:

Min-Max Normalization, K-Means Clustering, Enhanced Harmony Search, Gene expression, Feature selection, Classification

1. INTRODUCTION

Glioma begins from glial cells and neuronal precursors, and constitutes 80% of all malignant primary brain and CNS (Central Nervous System tumors). Glioblastoma represent about 15% of all major brain tumors [1]. They are malignant Grade IV tumors, where a large segment of tumor cells is reproducing and separating at any given time. The tumor is principally made up of abnormal astrocytic cells, but also holds a mix of different cell types (including blood vessels) and areas of dead cells (necrosis). Glioblastoma are commonly originated in the cerebral hemispheres of the brain, but can be found anyplace in the brain. Glioblastoma are a little more common in men than in women. In general, these tumors tend to be slower in increasing initially, but can progressively develop into aggressive. IDH (isocitrate dehydrognase) mutant glioblastoma account for approximately 10% of all glioblastoma. Glioblastoma are usually diagnosed as either IDH-wildtype or IDH-mutant [21]. IDH-wildtype glioblastoma is more common, tend to be more aggressive, and have poorer prognosis than IDH-mutant glioblastoma. It is exceedingly rare for glioblastoma [2] to extend outside of the brain.

Numerous studies have focused on the genetics of this tumor to further dissect the underlying mechanisms and to contribute to a better prognosis. Over the last decade several genetic lesions including TP53 and PTEN mutations [3] have been identified in glioblastoma tissue. There is strong epidemiologic evidence of family clustering of this tumor. Familial gliomas occur in approximately 5% of all glioma cases, the majority of which is associated with neoplastic syndromes like the Li-Fraumeni syndrome [4] and neurofibromatosis type 1. The genetic mutation happens in the normal cells lead to tumor. A gene is the basic physical and functional unit of heredity. Genes are made up of DNA. DNA, short for deoxyribonucleic acid, is the molecule that contains the genetic code of organisms [5]. DNA is in each cell in the organism and tells cells what proteins to make. Gene expression refers to the process of producing a protein, the final product of DNA. Genetic information is transcribed into mRNA and translated by the amino acid sequence of the protein [6]. Translated genetic information catalyzes biological reactions or forms of specific structures and is expressed in cells and individuals. During this process, when a gene becomes abnormal, it creates the wrong protein and mutations take place.

Glioblastoma due to genetic reasons is caused when one such process occurs. Mutant genes [7] that cause disease can be identified through special genetic [8] tests. These state-of-the-art tests enable early diagnosis, treatment, and active prevention, but they are expensive and suffer from the disadvantage that the patient has to wait for approximately a month for the test results. In addition, it is not easy to identify the mutant gene using these tests as [9] the probability of having a gene that causes Glioblastoma is 3–5%, [10] considering the total number of genes that make up the human body. It is difficult to choose a small number of genes compared to the high cost of genetic testing and the total number of genes.

In this proposed method the aforementioned difficulties can be overcome with the help of diagnosing genetic Glioblastoma. This study proposes the following feature selection method. First, candidate genes are selected for distribution between normal and abnormal classes using the generalized fisher score. Based on the data selected as a subset, K-means clustering is performed and representative genes for each cluster are found. Subsequently, using the harmony search (HS) algorithm, representative genes are searched for the optimal combination, which leads to high classification accuracy by using only a few genes.

2. RELATED WORK

The prediction of genetic diseases can be identified by the DNA information [11]. The complexity of diagnosis increased, because of the huge quantity of data or genetic mutations. In recent studies, with the progress made in the field of artificial intelligence, research of predicting diseases by means of biological data [12] has been dynamically conducted.

This TCGA dataset contains summing up of data visualizations and clinical data from a broad sampling of 592 glioblastoma multiformes. The data was gathered as part of the PanCancer Atlas initiative, which aims to reply big, overarching questions in relation to cancer by examining the full set of tumors characterized in the robust TCGA dataset. The clinical data includes mutation count, in order about mutated genes, patient demographics, disease status, tumor typing, and chromosomal gain or loss. The data set also includes copy-number segment data downloadable as .seg files and viewable via the Integrative Genomics Viewer.

In the above study, data were analyzed by means of random ensembles, and a support vector machine was used as a classifier to calculate MRI brain tumor type based on cancer gene information [13]. TCGA portal for cancer genomics is used to present the cancer data. R/MATLAB is obtainable with this tool for processing the gene information. The ability to supervise the evolution of the glioma genome through a minimally insidious technique [1] could move forward the clinical improvement and use of genotype-directed therapies for glioma, one of the most antagonistic human cancers.

There is also a study on feature selection using K-means clustering, wherein classification presentation was compared using known methods, such as mRMR, SVM-RFE, HSIC-LASSO, Clustering + mRNR, Clustering + SVM-RFE, and Clustering + HSICLASSO.

3. MATERIALS AND METHODS

The TCGA-cBioPortal for Cancer genomics provides researchers and physicians an insight of large-scale genomics data sets [2], helping out these persons to create and select better treatments for the patients. Glioma (MSKCC, Clin Cancer Res 2019) data set has been taken for processing in this study. 924 patient details are available in this dataset with 1004 samples. The number of samples per patient and mutation count is taken differently for patients. The sample types are of primary and recurrence type.

Totally 2000 GBM gene types are taken from 58 people for the purpose of classification from the data set TCGA-cBioPortal. The data used in this experiment are of types: DNA copy-number alterations, mRNA and microRNA expression, DNA Methylation, protein abundance and phospoprotein abundance [22]. The Fig.1 shows the steps proposed in this study. The parameters used in this study and their threshold values taken for this experiment is described in the process step by step.

3.1 MIN-MAX NORMALIZATION

Normalization is the course of action of scaling the attribute into a smaller particular range [14]. The range of the attribute may be defined in the range of -1.0 to 1.0 or 0.0 to 1.0. This procedure is particularly useful for classification algorithms relating neural networks, or distance measurements such as clustering and adjacent neighbor classification. Normalizing the input values for each attribute calculated in the training samples will help pace up the learning phase. The common techniques used for normalization is Min-Max normalization, Z-Score normalization and Decimal scaling.

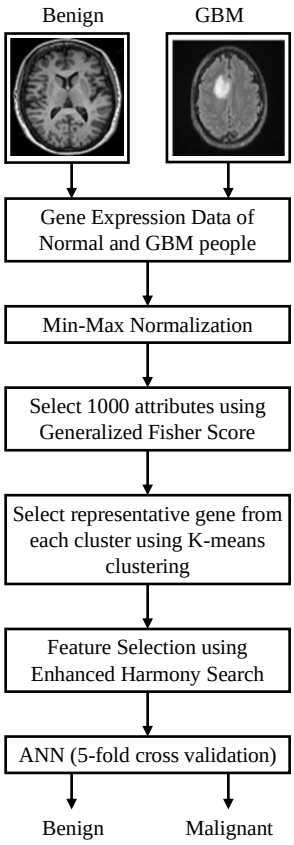

Fig.1. Schematic diagram of Proposed Method

Min-Max normalization performs a linear alteration on the original data. *MinA* and *MaxA* are the minimum and maximum values of an attribute A [15]. Min-Max normalization maps a value of *A* to *v*′ in the range by computing *new_MinA*, *new_MaxA*. The threshold value for the Min-Max range is fixed as [0, 1]

$$V' = \frac{V - MinA}{MaxA - MinA}\left(new\_MaxA - new\_MinA\right) + new\_MinA \quad (1)$$

Table.1. Min-Max Normalized values of 10 patients with respect to Attribute-1

Patient Number	Value of Attribute-1	Min-Max Normalized Value
1	3735.7501	−1.031397365
2	3332.3087	−1.223890725
3	6356.4587	−0.248737577
4	2410.3552	−1.45667784
5	4763.2275	−0.763844707
6	4872.1662	−0.660730665
58	9712.3725	0.67785507
59	7730.625	−0.092196709
61	6434.6225	−0.252561151
62	7572.01	0.147503275

The normalization is completed by substituting the innovative genetic information value into Table.1. The Table.1 shows the

values used with Min-Max normalization of Attribute-1 of each genetic information values for every patient [16]. The count of patients included in the actual experiment was 58, but Table.1 only shows, as an example, the value of Attribute 1 for 10 different patients taken from the dataset. The average of Attribute 1 gene information of 58 was 0.34567. The average of Attribute 1 is subtracted from the patient's genetic information (MinA of Attribute-1) value and divided by the difference of Maximum value of Attribute A and minimum value of Attribute A of the gene information [17]. As a result of this, a normalized number is obtained using the threshold value using within the specified threshold range of [0,1], as listed in the Table.1, which applies to all data.

3.2 GENERALIZED FISHER SCORE

A generalized Fisher score for feature selection. Rather than selecting each feature alone the proposed method selects a subset of features at the same time. It aims to locate a subset of features, which maximize the lower bound of conventional Fisher score [18]. It is able to consider the blend of features, and eliminate the unnecessary features. The resulting feature selection problem is a mixed integer programming,

This is further reformulated as a quadratically controlled linear programming (QCLP). It can be solved by cutting plane algorithm, in each iteration of which a various kernel learning problem [19] is solved by multivariate ridge regression and projected gradient descent alternatively.

$$F(W,p)=\mathrm{tr}\{(W^T\mathrm{diag}(p)S_b\mathrm{diag}(p)W)(W^T\mathrm{diag}(p)(S_t+\gamma I)\mathrm{diag}(p)W)^{-1}\},$$
$$\text{Subject to: } p\in\{0;1\}^d;\ p^T1=m\text{: where } W\in R^{d\times c} \quad (2)$$

3.3 K-MEANS CLUSTERING

The k-means algorithm takes the input parameter *k*, and partitions a set of *n* objects into *k* clusters. So, the resulting intra-cluster similarity is high but the inter-cluster similarity is low [28]. Cluster similarity is measured in regard to the mean value of the objects in a cluster. This can be viewed as the cluster's center of gravity.

Procedure of K-Means algorithm

Step 1: It randomly selects *k* of the objects, each of which initially represents a cluster mean or center.

Step 2: For each of remaining objects, an object is assigned to the cluster to which it is the most alike.

Step 3: Based on the distance connecting the object and the cluster mean, it then computes the new mean in favor of each cluster.

Step 4: This process iterates continually until Principle Function converges. The squared error criterion is defined as:

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \quad (3)$$

where, *E* is the summation of squared errors for all the objects in the database. *p* is the position in space representing a given object and m_i is the mean of cluster C_i. The criterion tries to make the resulting *k* clusters as dense and as separate as possible. This method is relatively scalable and capable in processing bigger data sets [23]. The computational complication of the algorithm is O(*nkt*), where *n* is the total quantity of objects, *k* is the number of clusters and *t* is the number of iterations.

In this study, the overall number of clusters was set to 20. The cluster consists of samples divided for the classification of MRI brain image [24]. Using all 1000 genes for feature selection, 20 representative genes were selected to account for variety. In each cluster, the gene whose information data were closest to the median value was chosen as the representative gene of the cluster [25]. The distance between the data and the median is calculated using the cosine distance. The cosine distance between the data points *u* and *v* can be calculated by the Eq.(4). The weights for each value is *u* and *v*. The computation of the cosine distance using a scipy, spatial and distance library.

$$W=1-\frac{u\cdot v}{\|u\|_2\|v\|_2} \quad (4)$$

3.4 ENHANCED HARMONIC SEARCH

Step 1: Parameters and Memory initialization of Harmony Memory

The first step in this harmony search is to initialize the variables and the harmony values to employ the harmony memory. The factor used in this algorithm has to be known for better understanding. As the working principle of HS algorithm [30] is analogous to an evolutionary algorithm, it can be compared well with a genetic algorithm. The genes are the fundamental elements of the chromosome in the genetic algorithm, is alike as the tones used in musical instruments. The tones are the basic element that constitutes the harmony vector array [31]. The Harmony memory size refers to the total number of harmonies in one harmony memory. They are randomly initialized in the preliminary stages of HS method implementation, and in the consecutive iterations the preceding harmony values are used for the follow up progression.

Random vectors (V_1, V_2,…,V_{HMS}) are generated, as many as Harmony memory size and assigned to the Harmony Matrix as below:

$$HM\text{-}Matrix=\begin{bmatrix} v_1^1 & v_2^1 & v_n^1 & f_x^1 \\ v_1^2 & v_2^2 & v_n^2 & f_x^2 \\ v_1^{hms} & v_2^{hms} & v_n^{hms} & f_x^{hms} \end{bmatrix} \quad (5)$$

Step 2: Creation of a new harmony

In this phase, ratio for grouping can be in tune and a new harmony can be created with wider variety of combinations. A set of harmonies created as many as HMS [32] is created in one harmony memory. From each memory harmony a distinctive harmony vector is randomly selected within the same location. The selected harmony vector becomes a novel harmony vector at the corresponding position in the harmony memory. Latest values at a location subsequent to each variable in the harmony are grouped together to create a new harmony. 1-HMCR is the probability of randomly initializing a harmony vector when creating the first harmony, in the succeeding stages a harmony is created and added to the harmony memory. The variant to the harmony vector is motivated by tuning [34] the pitch adjusting rate (PAR). PAR is used to achieve a various set of combinations.

- Generate new harmonies v' then for every element of v':

- With the probability value of HMCR (Harmony Memory Considering Rate: $0 \leq HMCR \leq 1$)
- Select a value from the HM such that

$$v_I' \leftarrow v_i^{(int(rand[0.1]XHMS)+1)} \quad (6)$$

- With the probability value of (1-*HMCR*), perform the uniform exploration between lower and upper bounds [35].

Step 3: Update Harmony Memory

In this phase, the newly occupied harmony vector is evaluated. The implication of the harmony is tested based on the threshold value of the harmony. If the new harmony generated in Step 2 is better, than the worst fit present in the harmony memory is eliminated, the new one vector is integrated in the harmony memory.

The HMCR and PAR variables can be updated [36] with the position update value and genetic mutation respectively. This will put off the algorithm from getting trapped in the local optimum. Here, V_{Bestj} and V_{worstj} are the best and worst V_i in HM, respectively, based on the objective function $f(x)$; vU and vL are the upper and lower bounds of the objective function, respectively; and rand .0 ≈1 is a random value between 0 and 1. Another modification on the original HS [37] is when the worst value is updated with the new V_j even if the new value is not better than the worst one.

Step 4: Repeating Step 2 and Step 3

Steps 2 and 3 are repeated as many times as the particular iteration. With each iteration, [38] the harmony with the lowest fitness is detached, and thus, a range of combinations are generated with the harmony of high fitness.

However, a new method of feature selection by modifying the existing HS. The related pseudocode is shown in Algorithm 1:

Algorithm 1

Step 1: Initialize the Parameters *PAR*, *BDR*, *HMS* and *HMCR*.

Step 2: Set i=0 [initialization of iterative variable]

Step 3: Assign Boolean values 0 and 1 for Initial Harmony

Step 4: *BDR*=*HMS**0.1// Assignment of upper and lower bound area

Step 5: Do

Step 6: Generate the initial harmony; i++

Step 7: While ($i \leq HMS$)

Step 8: While(j=1:N) //upper area harmony search

Step 9: v_{new} = Random Selection in the range v1j to v(BDR)j

Step 10: Generation of a new harmony (v_{new})

Step 11: If(Random(0,1)<*HMCR*) then //Lower area harmony search

Step 12: For (j=1:N) then v_{new} = Randomly select in the range of $v_{(BDR)1j}$ to $v_{(HMS)j}$

Step 13: If (Random(0,1)< *PAR*) then {|v_{new}= v_{new}-1|}

Step 14: End if;

Step 15: End For

Step 16: Generate new harmony (v_{new})

Step 17: Else

Step 18: Generate a harmony in Random manner

Step 19: End if

Step 20: For j = 1 to Each Dimension D // Memory updating by elimination of local optimum

Step 21: $v_R = 2v_j^{Best} - v_j^{Worst}$

Step 22: If(v_R>v_u)

Step 23: v_R=v_u

Step 24: else if (v_R<v_l)s

Step 25: v_R=v_1

Step 26: End if

Step 27: $v_j' = v_j^{Worst} + rand(0 \approx 1) * v_R - v_j^{Worst}$

Step 28: If (rand (0≈1)≤*PMR*)

Step 29: v_j' = v_l+rand(0≈1)*(v_u-v_l)

Step 30: End if

Step 31: Update the memory

Step 32: Update i = i+1

Step 33: Until (i≤max_iteration)

Step 34: Derive the Best Harmony.

Algorithm 2

Step 1: Initializing Variable and Harmony

Primarily the harmony vector is initialized with the values 0 and 1, to create a combination of 20 representative genes. Value 0 in the representative gene information indicates, it is not used as a feature for classification principle [39]. Value 1 in the gene information represents that it is used as a feature for classification task. The parameter value for HMCR is assigned as 0.9 and PAR is assigned as 0.1, and the number of iterations (i=300). HMS parameter value is set to 30.

Step 2: Formation of New Harmony memory and division of harmony memory

This step is a superior part of the existing HS for this work [28]. The creation of New Harmony memory follows the identical steps as in the existing HS algorithm. The research was conducted by separating the harmony memory into two distinct areas, as shown in Fig.2.

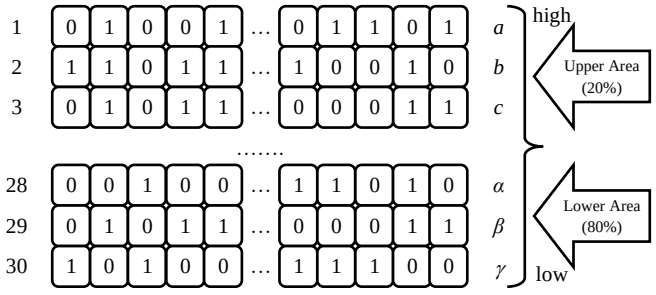

Fig.2. Harmony Memory Partition

The upper most area holds the harmonies having the fitness of the top 20% within one harmony memory area. The parameters HMCR and PAR are not taken into considerations in this step [24] The combination of higher fit can be found, when the combination is recombined within harmony of the upper most area. The following process is the formation of new harmonies. In the lower

area, new harmonies are created by the use of existing HS algorithm, which is by using HMCR and PAR parameters [20].

Step 3: Updating Harmony Memory

The fit is calculated based on each harmony value and is arranged in the order of harmony with the higher-level fitness. Two old harmonies with the lowest fitness value are assigned as per the Fig.3 and removed to match the size of the HMS that was initially specified [19].

The local optima search is not applied in this approach. By successive re-assignment of fitting the best and worst harmony in the process the local optima search is not used and avoids the search space significantly in the HMS. The generated worst harmonies are then eliminated from the Memory [23].

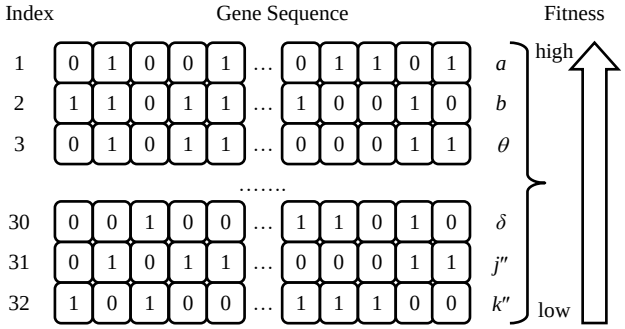

Fig.3. Poor Harmony Elimination

Step 4: Repeating Step 2 and Step 3

Based on the number of iterations assumed the step 2 and step 3 are repeated continuously. The upper memory region holds the harmonies with a higher degree of fitness inside the combination and with higher appropriateness [14]. The lower region takes the advantages of the exiting HS with the combinations according to the variety. As the count of iterations increases, the accuracy of classification task is advanced by storing two areas within one harmony as a text file. Classification accuracy changes as the iteration progress increases [13].

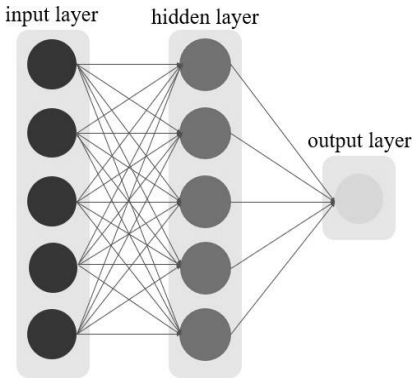

Fig.4. Architecture Diagram of ANN

4. CLASSIFICATION AND VALIDATION PROCESS

In this study, an artificial neural network has been used as a classifier [35]. It is a network that abstracts the working principle of human brain. The Fig.4 shows the architecture of the ANN used in this study. 5 nodes are assumed for input and hidden layers. The output layer consists of a single node, and the activation function taken here is sigmoid function.

A 5-Fold cross validation technique is functional for the experimental verification process [37]. The reliability of the data verification can be enhanced by using all the data as a test set at least once.

The Fig.5 shows the method of training and testing data by means of 5-fold cross validation. The significant feature selected through the enhanced harmony search algorithm is verified through this 5-Fold cross validation approach [8].

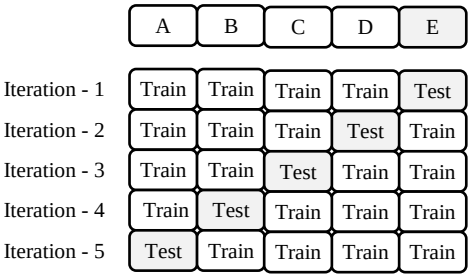

Fig.5. 5-Fold Cross Validation Technique

4.1 RESULTS

In this study, a total of 2000 genes are chosen through the generalized fisher score out of which 1000 candidate genes are [32] selected and 20 clusters are bent by using the K-Means clustering technique. The Scikit environment decides the best possible number of clusters using the inertia value. The inertial value vary depends upon the number clusters used in this study. For lower inertia value the distance between the cluster and the centroid is closer [34]. Assigning minimal inertia value, creates a higher degree of aggregation of the data in the cluster to be evaluated. But too many clusters formation will increase the classification error rate.

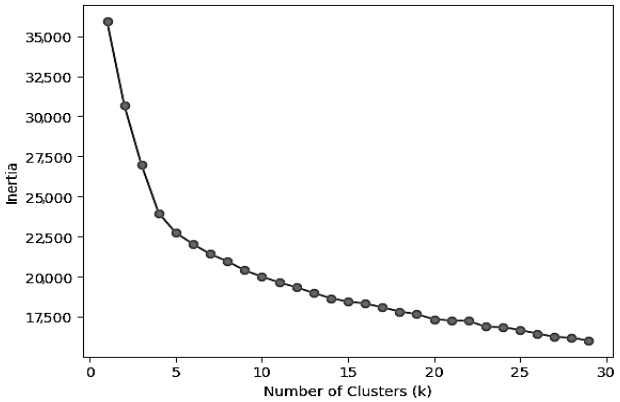

Fig.6. Inertia Value with respect to Number of clusters

There were 104 genes were common between Lower Grade Glioma (LGG) and Glioblastoma (GBM). Patient survival rate is based on sequencing the gene data obtained from The Cancer Genome Atlas, which contains the gene expression for analysis [35]. The table represents a gene value according to a patient's attribute, and each column is used to represents a patient's gene information value for each attribute. The enhanced harmony search method selects 10 genes from 20 representative clusters.

Table.2. Representative genes used in Enhanced Harmony Search

	352	455	742	720	1635	992	936	1897	1515	318	1244	1170	1177	737	640	482	109	980	43	33	AA
1	-0.199	0.527	0.102	-0.068	-0.668	-0.030	-0.549	-0.663	0.137	-0.053	-0.364	-0.653	-0.522	-0.310	-0.218	-0.884	-0.842	0.479	-0.223	-0.263	0
2	-0.767	0.440	0.576	0.372	0.997	0.424	-0.251	0.449	-0.020	-0.643	1.883	-0.334	-0.115	1.279	0.380	-0.443	0.110	0.758	-0.726	0.006	1
3	3.057	-0.116	1.235	-0.524	-0.814	-0.992	-0.766	-0.811	-1.272	-1.290	0.724	-1.191	-0.532	-0.825	-0.881	-1.125	-1.164	0.258	-1.318	-1.088	0
4	1.270	-0.090	0.874	-0.238	-0.594	-0.813	-0.718	-0.590	-0.522	-0.931	-0.334	-0.989	-0.699	0.133	-0.422	-0.876	-1.129	0.364	-1.031	-0.791	1
5	-0.065	-0.332	-0.681	-0.845	-0.651	-0.980	-0.561	-0.422	-1.170	-0.622	-0.362	-0.525	-0.347	-1.021	-0.384	-0.381	-0.358	0.352	0.190	-0.638	0
6	-0.261	-0.496	-1.119	-1.252	-0.259	-1.014	-0.766	0.075	-1.268	-0.647	-0.759	-0.989	-0.512	-0.799	-0.914	-0.716	-1.186	-0.036	-0.935	-0.624	1
7	-0.673	0.989	-0.134	-0.029	-0.747	-0.307	-0.373	-0.669	-0.886	-0.681	-0.862	-0.898	-0.354	-1.038	-0.270	-0.553	-0.534	-0.019	-0.558	-0.732	0
8	-0.631	-0.136	-0.842	-0.478	-0.382	-0.712	-0.313	-0.681	-1.254	-0.377	-0.301	-1.015	-0.187	-0.365	-0.625	-0.468	-1.059	1.036	-1.044	-0.364	1
9	-1.048	2.248	1.639	0.170	-0.871	0.558	0.897	-0.558	1.476	0.814	-0.582	-0.899	-0.246	-0.336	0.910	0.984	-1.053	0.220	1.712	-1.674	0
10	-0.255	0.390	-0.129	-0.483	0.596	-0.704	-0.531	0.529	-1.063	-0.563	-0.381	0.111	0.994	-0.513	-0.103	-0.270	0.537	0.222	-0.510	0.417	1
11	0.464	3.183	-0.375	2.554	0.339	3.096	0.829	-0.247	-0.784	1.061	-0.122	3.411	3.189	-0.824	5.506	2.061	3.051	1.605	1.604	-0.081	0
12	1.064	0.815	0.813	0.794	0.404	0.729	-0.334	0.697	0.276	0.003	1.397	-0.276	0.250	1.467	0.566	-0.318	0.103	0.047	0.012	-0.368	1
13	-0.360	-0.378	-0.528	-0.017	-0.219	-0.688	0.004	-0.377	-0.995	-0.221	-0.542	-0.218	-0.524	-0.737	-0.543	-0.401	0.213	0.352	-0.085	-0.248	0
14	0.673	-0.634	-0.589	-0.766	-0.050	-0.900	-0.528	-0.162	-0.768	-0.472	-0.540	-0.873	-0.892	-0.567	-0.712	-0.739	-0.446	0.025	-0.467	-0.542	1
15	0.150	-0.489	-0.064	-0.639	-0.418	-0.699	-0.148	-0.474	-0.391	0.307	-0.194	-0.453	-0.246	-0.210	-0.613	-0.181	-0.337	2.824	0.460	0.323	0
16	0.168	-0.764	-0.611	-1.221	-0.585	-1.129	-0.276	-0.830	-0.889	0.106	-0.873	-0.873	-0.895	-1.060	-0.915	-0.324	-0.033	0.417	0.007	-1.128	1
17	-0.572	-0.846	-1.270	-0.383	-0.892	-0.384	-0.561	-0.644	-1.223	-0.716	-0.865	-0.988	-1.132	-0.855	-0.458	-0.612	-0.995	-1.021	0.600	-0.973	0
18	-0.966	-0.674	-1.212	-0.783	-0.089	-0.472	-0.844	-0.552	-1.284	-1.144	-0.981	-1.070	-1.219	-0.961	-0.443	-0.849	-1.189	-0.363	-1.026	-0.651	1
19	-1.044	-0.209	-0.612	-0.275	-0.954	-0.638	-0.248	-0.690	-0.737	-0.134	-0.743	-0.727	-0.908	-0.885	-0.045	-0.159	-1.249	-0.340	0.238	-0.913	0
20	-0.253	0.153	-1.222	-0.872	0.436	-0.949	-0.573	0.030	-1.306	-0.552	-0.798	-0.637	-0.138	-0.991	-0.675	-0.600	-0.920	-0.048	-0.388	-0.181	1
21	-0.922	1.024	0.186	0.133	-0.526	0.615	-0.050	-0.654	0.186	-1.465	-0.747	0.142	0.122	-0.245	-0.048	-0.266	-0.191	0.711	-0.282	0.592	0
22	1.192	0.542	0.628	0.228	1.424	0.257	-0.649	0.190	0.123	-0.481	0.245	0.254	0.437	0.526	0.011	-0.802	0.481	1.496	-0.791	1.930	1
23	-1.000	-1.296	-1.332	-0.997	-0.799	-1.045	-0.171	-0.697	-1.193	0.008	-0.921	-0.810	-0.779	-1.037	-0.530	0.194	-0.151	0.260	0.255	-0.684	0
24	-1.125	-1.320	-1.518	-1.587	-0.600	-1.217	-1.035	-0.802	-1.393	-1.385	-1.047	-1.192	-1.257	-1.050	-0.962	-1.055	-1.268	-1.024	-1.420	-1.253	1
25	2.153	-0.380	0.413	0.058	-0.611	-0.269	-0.015	-0.523	0.559	0.724	-0.772	-0.262	0.089	-0.560	-0.330	0.533	0.394	-0.840	1.521	-0.046	0
26	-0.815	-0.333	-0.528	-0.208	-0.780	-0.686	-0.797	-0.694	0.017	-0.812	-0.733	0.052	-0.175	-0.894	-0.522	-0.670	0.571	-1.262	-0.881	-0.901	0
27	-0.774	-0.602	-0.611	-0.538	-0.454	-0.242	-0.747	-0.582	0.024	-0.715	-0.688	-0.091	-0.197	-0.772	-0.584	-0.693	0.130	-1.323	-0.749	-0.874	0
28	-0.911	0.517	1.098	0.555	-0.697	0.265	0.232	-0.552	2.166	1.076	-0.039	0.318	0.772	0.431	0.481	0.589	1.361	-0.521	0.926	-0.210	0
29	-0.572	2.755	3.127	1.969	-0.009	1.186	-0.048	-0.189	2.991	0.420	0.357	2.058	3.010	-0.147	1.005	-0.139	2.787	-0.616	-0.109	0.457	0
30	-0.916	1.660	1.195	2.319	-0.494	1.794	0.405	-0.311	1.471	0.468	-0.325	1.044	1.833	0.132	0.950	0.386	2.165	-0.355	0.609	-0.090	0
31	-1.117	0.883	1.863	0.298	-0.360	0.268	0.357	0.390	2.500	0.958	4.194	0.275	1.211	1.057	-0.093	0.249	0.481	-0.820	0.326	0.544	0
32	-0.983	-0.196	-1.357	-0.414	0.415	0.580	-0.527	0.111	-0.944	-0.746	-0.481	0.141	-0.281	-1.009	-0.201	-0.281	-0.381	-1.155	-0.517	-0.871	0
33	-0.496	-0.013	-1.179	-0.239	-0.538	-0.336	-0.380	-0.608	-0.944	-0.118	-1.015	-0.304	-0.089	-1.076	-0.438	0.361	-0.355	-1.286	0.454	-0.465	0
34	0.984	0.135	0.854	0.138	-0.442	0.036	0.705	-0.201	0.881	1.006	0.334	-0.173	0.436	0.347	-0.190	0.142	0.371	-0.490	0.348	-0.397	0
35	-0.483	-0.339	-0.622	-0.129	-0.679	0.183	-0.300	-0.567	-0.140	-0.404	-0.225	-0.325	-0.600	-0.852	-0.294	-0.336	-1.043	-1.059	-0.886	-0.801	0
36	-0.728	-0.870	-1.087	-0.855	-0.817	0.097	0.564	-0.548	-0.118	0.341	-0.681	-0.228	-0.814	-0.896	-0.607	0.606	-0.419	-0.850	0.591	-1.144	0

The selected genes are as follows: CTSZ (cathepsin Z), EFEMP2 (EGF-containing fibulin-like extracellular matrix protein2), ITGA5 (integrin alpha-5), KDELR2 (KDEL Endoplasmic Reticulum Protein Retention Receptor 2), MDK (midkine), MICALL2 (Junctional Rab-13 functional protein), MAP 2 K3 (Mitogen-Activated Protein Kinase Kinase-3), PLAUR (Plasminogen Activator, Urokinase Receptor) SERPINE1 (endothelial plasminogen activator inhibitor), and SOCS3 (Suppressor of cytokine signaling-3). The classification accuracy [41] using the ANN architecture provides 96.67 %. Each attribute is significantly related to GBM type tumor, and the evidence for this is supported by research studies.

5. COMPARITIVE ANALYSIS

Numerous researchers have experimented with a set of classification algorithms using the GBM tumor data provided by the Cancer Genome Atlas. The classification accuracies with respect to the number of gene selected [46] for different studies are shown using table. The number of genes selected and the classification accuracy of each research study differs in their representative accuracy parameter. There are studies that provide the classification accuracy without taking into the account of feature selection phase. In recent studies, many researchers have used support vector machine (SVM), random forest (RF) and LogitBoot for 10-cross validation on the dataset provided by the

cancer genome atlas. Furthermore, the classification accuracy has been derived through feature selection by using Chameleon algorithm and supervised group Lasso method.

Table.3. Performance Comparison

Selected Genes	Method	Accuracy (%)
2000	Random Forest	85.24
2000	SVM	84.42
2000	Two-Way Clustering	87.76
2000	LogitBoot	86.67
5	Chameleon	86.23
22	Lasso (Supervised Group)	86.64
10	MM-FS-KM-EHS (Proposed)	96.67

The projected method achieved the highest level of accuracy, when compared with other methods. The achieved accuracy is compared with the studies that consider the feature selection as a step or no feature selection involved. The Chameleon algorithm is able to achieve significant accuracy level by taking minimum number of genes for its classification task [44]. Still, the projected method achieved better accuracy compared with the Chameleon algorithm.

6. CONCLUSION

The classification procedure in this study uses GBM gene information. Min-Max normalization is used to preprocess and normalize the gene information in the initial step. The Generalized fisher score method is used to eliminate the redundant genes to provide optimal set of gene information. The K-Means clustering method selects the representative genes from each cluster. The enhanced feature selection using enhanced harmony search method is used for critical feature selection process.

The proposed improved feature selection process is done through enhanced harmony search method derived from the original HS algorithm, which retains higher accuracy and improves the performance of classification by applying different combinations of the model. This study showed a promising classification performance of 96.67% with only 10 genes selected using the proposed method. The selected attributes are: attribute1635, attribute936, attribute1897, attribute1515, attribute1170, attribute1177, attribute737, attribute43, attribute 33 and attribute1244. This method takes only minimum number of genetic test information, which is cost-effective. Furthermore, the outcome of the study will contribute significantly in the prediction of not only the GBM gene. This method can be applied efficiently with other gene causing diseases. Heredity based colon cancer can also be predicted by using this genetic testing-based study. The likelihood confirmation through gene testing for people related with the family history of cancer related diseases is important. It improves the prediction accuracy of the people, who likely to develop tumor and takes precautionary medical assistance from the physician's advice.

In future, prediction of tumor can be found by using a minimum count of representative genes according to gene mutation. There is a possibility of conducting experiments on gene expression analysis in different ways. The analysis of gene information can be done by methods such as solo-atom tracing, time series data tracking. The analysis can be applied on non-genetic data like smoking, diet and exercise-based gene information. New Models can be developed based on objectivity and suitability of the genetic data from the proposed model.

REFERENCES

[1] Liang-Bo Wang, "Proteogenomic and Metabolomic Characterization of Human Glioblastoma", *Cancer Cell*, Vol. 39, No. 4, pp. 509-528, 2021.

[2] Christina A Clarke, "Multi-cancer Early Detection: A New Paradigm for Reducing Cancer-Specific and All-Cause Mortality", *Cancer Cell*, Vol 39, No. 4, pp. 447-448, 2021.

[3] Anne Le, "The Multifaceted Glioblastoma: From Genomic Alterations to Metabolic Adaptations", *The Heterogeneity of Cancer Metabolism*, Vol. 1311, pp. 59-76, 2021.

[4] K. Nabi and A. Le, "The Intratumoral Heterogeneity of Cancer Metabolism", *Advances in Experimental Medicine and Biology*, Vol. 1311, pp.1-12, 2021.

[5] S. Bose, C. Zhang and A. Le, "Glucose Metabolism in Cancer: The Warburg Effect and Beyond", *Advances in Experimental Medicine and Biology*, Vol. 1311, pp. 98-115, 2021.

[6] Luciano Garofeno, "Pathway-Based Classification of Glioblastoma Uncovers a Mitochondrial Subtype with Therapeutic Vulnerabilities", *Nature Cancer*, Vol. 2, pp. 141-156, 2021.

[7] Zeyu Wang, "The Adaptive Transition of Glioblastoma Stem Cells and its Implications on Treatments", *Signal Transduction and Targeted Therapy*, Vol. 6, pp. 124-135, 2021.

[8] Nguyen Quoc Khan, "Randiomics-based Machine Learning Model for Efficiently Classifying Transcriptome Subtypes in Glioblastoma Patients from MRI", *Computers in Biology and Medicine*, Vol. 132, pp. 1-12, 2021.

[9] Paul Minh Huy Tran, "Retrospective Validation of a 168-Gene Expression Signature for Glioma Classification on a Single Molecule Counting Platform", *Cancers*, Vol. 13, No. 3 pp. 1-13, 2021.

[10] T. Weiss and M. Weller, "Pathway-based Stratification of Glioblastoma", *Nature Reviews Neurology*, Vol. 17, pp. 263-264, 2021.

[11] Chen Ma, "Quantitative Integration of Radiomic and Genomic Data Improves Survival Prediction of Low-Grade Glioma Patients", *Mathematical Biosciences and Engineering*, Vol. 18, No. 1, pp. 727-744, 2021.

[12] J.W. Robinson and S. Tsavachidis, "Transcriptome-Wide Mendelian Randomization Study Prioritising Novel Tissue-Dependent Genes for Glioma Susceptibility", *Scientific Report*, Vol. 11, pp. 1-12, 2021.

[13] Chiwen Qu, "Improving Feature Selection Performance for Classification of Gene Expression Data using Harris Hawks Optimizer with Variable Neighborhood Learning", *Briefings in Bioinformatics*, Vol. 23, No. 1, pp. 1-12, 2021.

[14] Ben Brahim, "A Stable Feature Selection based on Instance Learning, Redundancy Elimination and Efficient Subsets Fusion", *Neural Computing and Applications*, Vol. 33, pp. 1221-1232, 2021.

[15] P.M.H. Tran and R. Bollag, "Comparative Analysis of Transcriptomic Profile, Histology, and IDH Mutation for Classification of Gliomas", *Scientific Report*, Vol. 10, pp. 1-12, 2020.

[16] D.N. Louis, "Impact-Now update 6: New Entity and Diagnostic Principle Recommendations of the Impact-Utrecht Meeting on Future CNS Tumor Classification and Grading", *Brain Pathology*, Vol. 30, pp. 844-856, 2020.

[17] K. Willemsma, W. Yip and C.E. Simmons, "Impact of Recurrence Score on Type and Duration of Chemotherapy in Breast Cancer", Current Oncology, Vol. 27, No. 2, pp. 86-92, 2020.

[18] J. Zheng, "Phenome-Wide Mendelian Randomization Mapping the Influence of the Plasma Proteome on Complex Diseases", Nature Genetics, Vol. 52, pp. 1122-1131, 2020.

[19] Guozhang Hu, "Prognostic Markers Identification in Glioma by Gene Expression Profile Analysis", *Journal of Computational Biology*, Vol. 27, No. 1, pp. 1-8, 2020.

[20] P. Rawla, T. Sunkara and A. Barsouk, "Epidemiology of Colorectal Cancer: Incidence, Mortality, Survival, and Risk Factors", *Gastroenterologiczny*, Vol. 14, No. 2, pp. 89-103, 2019.

[21] Zhendong Liu, "Construction of Lncrna-Associated ceRNA Networks to Identify Prognostic Lncrna Biomarkers for Glioblastoma", *Journal of Cellular Biochemistry*, Vol. 121, No. 7, pp. 3502-3515, 2020.

[22] Zahra Jabbarpour, "Effects of Human Placenta-Derived Mesenchymal Stem Cells with NK4 Gene Expression on Glioblastoma Multiforme Cell Lines", *Journal of Cellular Biochemistry*, Vol. 121, No. 2, pp. 1362-1373, 2020.

[23] Yulin Wang, "A Risk Signature with Four Autophagy-Related Genes for Predicting Survival of Glioblastoma Multiforme", *Journal of Cellular and Molecular Medicine*, Vol. 24, No. 7, pp. 3807-3821, 2020.

[24] G. Hoang, S. Udupa and A. Le, "Application of Metabolomics Technologies Toward Cancer Prognosis and Therapy", *International Review of Cell and Molecular Biology*, Vol. 347, pp. 191-223, 2019.

[25] C.M. Jackson and M. Lim, "Mechanisms of Immunotherapy Resistance: Lessons from Glioblastoma", *Nature Immunology*, Vol. 20, pp. 1100-1109, 2019.

[26] N. Wijethilake, M. Islam and H. Ren, "Radiogenomics Model for Overall Survival Prediction of Glioblastoma", *Medical and Biological Engineering and Computing*, Vol. **58**, No. 8, pp. 1767-1777, 2020.

[27] R Core Team, "A Language and Environment for Statistical Computing", Available at https://www.R-project.org/, Accessed at 2019.

[28] H. Marvi Khorasani and H. Usefi, "Feature Clustering Towards Gene Selection", *Proceedings of IEEE International Conference on Machine Learning and Applications*, pp. 16-19, 2019.

[29] J. Xie, Y. Wang and Z. Yu, "Colon Cancer Data Analysis by Chameleon Algorithm", *Health Information Science and Systems*, Vol. 7, pp. 1-8, 2019.

[30] G. Munoz-Gil and C. Manzo, "Single Trajectory Characterization via Machine Learning", *New Journal of Physics*, Vol. 22, pp. 1-12, 2019.

[31] L. Follia, "Integrative Analysis of Novel Metabolic Subtypes in Pancreatic Cancer Fosters New Prognostic Biomarkers", *Frontiers in Oncology*, Vol. **9**, pp. 115-123, 2019.

[32] Fan Wu, "Molecular Classification of *IDH*-Mutant Glioblastomas based on Gene Expression Profiles", *Carcinogenesis*, Vol. 40, No. 7, pp. 853-860, 2019.

[33] R.R. Agravat and M.S. Raval, "Prediction of Overall Survival of Brain Tumor Patients", *Proceedings of International Conference on Machine Learning*, pp. 1-12, 2019.

[34] B.A. Alves Martins and A.M.A. Martins, "Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research", *Frontiers in Oncology*, Vol. 9, pp. 1-12, 2019.

[35] G. Munoz Gil and M. Lewenstein, "Single Trajectory Characterization via Machine Learning", *New Journal of Physics*, Vol. 2019, pp. 1-22, 2019.

[36] C. Aaberg Jessan, "Co-Expression Of TIMP-1 And Its Cell Surface Binding Partner CD63 In Glioblastomas", *BMC Cancer*, Vol. 18, No. 1, pp. 270-287, 2018.

[37] R.X. Geng, "Identification of Core Biomarkers Associated with Outcome in Glioma: Evidence From Bioinformatics Analysis", *Disease Markers*, Vol. 2018, pp. 1-16, 2018.

[38] C.Y. Lin, S.T. Yang and T.I. Hsu, "Serum Amyloid A1 In Combination with Integrin Avβ3 Increases Glioblastoma Cells Mobility and Progression", *Molecular Oncology*, Vol. 12, No. 5, pp. 756-771, 2018.

[39] S. Zhong, B. Wu and H. Zhang, "Identification of Driver Genes and Key Pathways of Glioblastoma Shows JNJ-7706621 As A Novel Antiglioblastoma Drug", *World Neurosurgery*, Vol. 109, pp. 329-342, 2018.

[40] F. Bray, J. Ferlay and A. Jemal, "Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA: A Cancer Journal for Clinicians*, Vol. 68, pp. 394-424, 2018.

[41] Q. Yang and J.Y. Wang, "Proteomic Profiling of Antibody-Inducing Immunogens in Tumor Tissue Identifies PSMA1, LAP3, ANXA3, and Maspin as Colon Cancer Markers", *Oncotarget*, Vol. 9, pp. 3996-4019, 2018.

[42] J. Ferlay and F. Bray, "Global Cancer Observatory: Cancer Today; International Agency for Research on Cancer: Lyon, Available at https://gco.iarc.fr/today/, Accessed at 2018.

[43] F. Chao, D. Zhou and D. Lin, "Fuzzy Cerebellar Model Articulation Controller Network Optimization via Self-Adaptive Global Best Harmony Search Algorithm", *Soft Computing*, Vol. 22, pp. 3141-3153, 2018.

[44] G. Li, B. Zeng and L. Gao, "A New AGV Scheduling Algorithm based on Harmony Search for Material Transfer in a Real-World Manufacturing System", *Advances in Mechanical Engineering*, Vol. 10, pp. 1-13, 2018.

[45] H. Ouyang, L. Gao and S. Li, "Amended Harmony Search Algorithm with Perturbation Strategy for Large-Scale System Reliability Problems", *Applied Intelligence*, Vol. 48, pp. 3863-3888, 2018.

[46] A. Sadollah, H. Sayyaadi and J.H. Kim, "Mine Blast Harmony Search: A New Hybrid Optimization Method for Improving Exploration and Exploitation Capabilities", *Applied Soft Computing*, Vol. 68, pp. 548-564, 2018.