Received February 28, 2020, accepted March 8, 2020, date of publication March 13, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980728

A Novel Approach to Improving Brain Image Classification Using Mutual Information-Accelerated Singular Value Decomposition

ZAHRAA A. AL-SAFFAR[1,2] AND TÜLAY YILDIRIM[2], (Member, IEEE)
[1]Department of Biomedical Engineering, Al-Khwarizmi College of Engineering, University of Baghdad, Baghdad 10071, Iraq
[2]Department of Electronics and Communications Engineering, Yildiz Technical University, 34220 Istanbul, Turkey

Corresponding author: Zahraa A. Al-Saffar (znzs_0507@kecbu.uobaghdad.edu.iq)

ABSTRACT Brain image classification is one of the most useful and widely needed processes in the medical system, and it is a highly challenging field. This paper presents a new method for selecting a significant subset of features as the input to the classifier, called mutual information-accelerated singular value decomposition (MI-ASVD). This novel algorithm is exploited to design an intelligent system for classifying MRI brain images into three classes: healthy, high-grade glioma, and low-grade glioma. The proposed system has six stages: pre-processing, clustering, tumour localization, feature extraction, MI-ASVD and classification. First, the MR images are smoothed by using enhancement techniques such as Gaussian kernel filters. Then, local difference in intensity-means (LDI-Means) clustering is employed to segment and detect suspicious regions. The grey-level run-length matrix (GLRLM), texture, and colour intensity features are used for tumour feature extraction. Later, a special method including a summation of feature selection and dimensionality reduction, MI-ASVD, is applied to select the most useful features for the classification process. Finally, the simplified residual neural network technique is implemented to classify the MR brain images. Using MI-ASVD provided accurate and more efficacious results in classification compared with the original feature space and with two other standard dimensionality reduction methods, principal component analysis (PCA) and singular value decomposition (SVD). It achieved a classification accuracy of 94.91%, which is better than the two state-of-the-art techniques as well as methods from similar published studies.

INDEX TERMS Brain image classification, clustering, image processing, machine learning, mutual information, PCA, residual neural network (RNN), SVD.

I. INTRODUCTION

The main purpose of brain tumour classification is to correctly classify MR images to detect which type of tumour is affecting a patient. Glioma is the brain tumour with the highest death rate and occurrence rate. These neoplasms can be graded as low-grade gliomas (LGGs) and high-grade gliomas (HGGs) according to their aggressive and infiltrative characteristics [1]. The treatment of a brain tumour depends on the tumour size, its type, and its growth stage. In the healthcare industry, doctors detect the presence of a brain tumour with the help of various medical imaging techniques [2].

Medical imaging can be defined as a technique used to create images for clinical research, diagnosis, and treatment. Magnetic resonance imaging (MRI) is an imaging technique that uses magnetic fields to capture images. It can provide reproducible, non-invasive, and quantitative measurements of tissue, including structural, anatomical, and functional information [3]. Because of its outstanding soft tissue contrast and detailed resolution, MRI is the most commonly utilized modality for brain tumour growth imaging and location detection. MR brain image classification and tumour detection are still based mostly on direct human inspection of these images.

The associate editor coordinating the review of this manuscript and approving it for publication was Charith Abhayaratne.

This visual evaluation and examination by radiologists is subjective by its nature, and it is time-consuming and prone to errors and omissions [4]. Therefore, to enhance physicians' diagnostic capabilities and reduce the time required, a medical decision-making system for automated brain tumour detection and classification of MR images has been developed.

In machine learning, the classification process is defined as a supervised learning task that infers a link between features (characteristics of the dataset) and class labels.

The classification of high-dimensional data is based on these extracted features [5].

However, a large number of features often leads to overfitting, high computational complexity, and low interpretability of the final model [6]. For better results, the medical decision-making field has begun to use data mining techniques to detect the presence of such tumours. Hence, physicians can use a brain tumour detection system as a second opinion in addition to their own view in finding the right diagnosis and treatment of brain tumours [1]. In the data mining field, the quality of input data determines the quality of the output, e.g., accuracy. The input data to any machine learning algorithm is approximately expressed by features showing different properties of the problem. Therefore, the quality of the feature space is key in solving an image analysis problem [7].

The objective of this paper is to develop an intelligent system that can classify MR brain images more effectively than existing models into three classes: healthy, high-grade glioma (HGG) and low-grade glioma (LGG). To achieve this aim, a novel algorithm named MI-ASVD has been used, which is a combination of two techniques: mutual information (MI) as a feature selection method and singular value decomposition (SVD) as a dimensionality reduction method. This algorithm has been considered a pre-step for classification process to improve the classifier performance.

The contributions of this paper can be summarized as follows:

- A clustering algorithm based on intensity, named LDI-Means (local difference intensity - means), is used to perform segmentation.
- A feature selection method based on MI is used to re-rank the extracted features, which will help the ordinary SVD technique to work automatically. Hence, there is no need for the user to specify the number of dimensions that is needed to reduce the features; this novel method is named MI-ASVD.
- A simplified version of RNN is used to perform classification, as shown in Fig. 1.

In addition to the introduction mentioned above, this paper is structured as follows: Section (II) presents some of the related literature. Section (III) explains the background theory. Section (IV) describes the methodology. Section (V) contains the experimental results as well as the discussion. Finally, the conclusions are addressed in section VI.

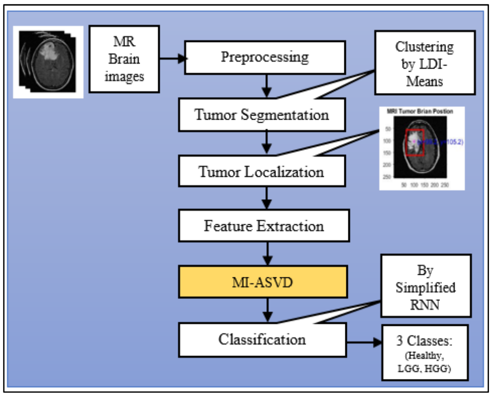

FIGURE 1. The block diagram of the proposed system. MI-ASVD is the novel part of this paper.

II. RELATED LITERATURE

Some of the data mining and machine learning approaches used to perform feature selection and classification are explained below:

Battiti [8] introduced a new way to perform feature selection by using mutual information theory, which is called MIFS. This algorithm includes calculating the relationships among each feature and all other features and among each feature and all classes. Then, another useful subset of features is formed by selecting a feature that gives a large amount of information about the class label. Peng *et al.* [9] presented another algorithm, called mRMR, based on MI theory. It is used to minimize the redundancy and maximize the relevancy among features. Kumar *et al.* [10] presented a principal component analysis - artificial neural network (PCA-ANN) system to assist radiologists in multiclass brain tumour classification. It consists of four stages: gradient vector flow (GVF), feature extraction, feature reduction by using PCA and classification by using an ANN. Additionally, a multiclass brain tumour classification method using PCA-ANN was presented by Sachdeva *et al.* in [11]. Eight hundred and fifty-six ROIs were obtained by using the content-based active contour (CBAC). After the CBAC step, more than one experiment was performed in this study to check the classification accuracy by using an ANN with and without PCA. The experimental results showed an increase in the accuracy from 77 to 91%.

Corso *et al.* [12] proposed a new method. It combines two of the most effective approaches to segmentation. The first approach uses generative model-based techniques, and the second uses a graph-based affinities method. Then, the multi-level segmentation by weighted aggregation algorithm (SWA) is used. A study by Fang *et al.* [13] computed the value of MI by using the Kozachenko Leonenko information entropy estimation method to find robust features. Then, it factorized the obtained feature matrices to satisfy the input of the classification stage. The results showed that this improved method could successfully decrease the number of dimensions of multidimensional time series for clinical data. Hoque *et al.* [14] presented a feature selection method based on fuzzy MI with a nondominated solution. This study selects features according to the fuzzy mutual information between each feature and class as well as between each feature and the other features. Additionally, this study presents a modification of the k-nearest neighbour (KNN) classifier to classify instances based on distance. A DTI algorithm is used in a study by Jones *et al.* [15] to delineate tumour VOIs by using isotropic and anisotropic properties of the diffusion tensor. Then, D-SEG spectra are considered within each VOI. The classification of tumour type using D-SEG spectra is implemented using an SVM.

More than one experiment was performed in a study by Zacharaki *et al.* [16]. This study suggested using conventional MRI and echo-planar relative cerebral blood volume (rCBV) maps for distinguishing brain tumour types and the support vector machine recursive feature elimination (SVM-RFE) algorithm as a feature subset selection method. Later, multi-class classification is performed by using an SVM.

Automated detection and segmentation of abnormal brain tissue from fluid-attenuated inversion recovery (FLAIR) MRIs was performed in a study by Soltaninejad *et al.* [17]. The classification is performed with the super-pixel technique. Two classifiers are used: extremely randomized trees (ERT) and an SVM. Additionally, Soltaninejad *et al.* [18] proposed a novel super-voxel-based learning method. A set of features, including histograms of texton descriptors, are computed as a result of using a number of Gabor filters in different orientations and sizes, and first-order intensity statistical features are extracted. The extracted features are used as an input to the random forests (RF) classifier to perform classification into the categories of tumour core, oedema or healthy brain tissue.

Dong *et al.* [19] introduced an automatic brain tumour detection system using a U-Net-based deep convolutional neural network. In [20], Bakas *et al.* identified the best machine learning methods used to analyse brain tumour images of MRI scans from 2012 to 2018 for the international brain tumour segmentation (BRATS) challenge.

III. BACKGROUND THEORY

Applying one of the machine learning or data mining algorithms in any model may face a critical risk when dealing with high-dimensional data or a wide feature space as a result of a large number of input variables. In general, two possible techniques can be used to overcome this problem. First, feature-selection algorithms can be used to choose only the most relevant variables from the original dataset. The second is to use dimensionality-reduction algorithms that take advantage of the redundancy of the input data to calculate new variables to be a new subset of the data [14]. PCA and SVD are the most widely used algorithms to perform dimensionality reduction.

A. FEATURE SELECTION BASED ON MUTUAL INFORMATION THEORY

Mutual information is a concept rooted in information theory. It is a statistical method that calculates how much information each variable has about the other variables [8]. For example, if A and B are independent, then A includes no information about B, and their MI is zero. If A and B are the same, then all information carried by A is shared with B, and knowing A reveals nothing new about B; accordingly, the MI is the same as the information carried by A or B alone, which is called the entropy of A [7], [21], [22].

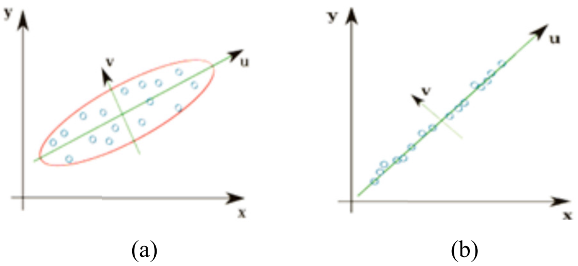

FIGURE 2. Principal component analysis in (a) the original feature space and (b) a reduced-dimension space [23].

B. DIMENSIONALITY REDUCTION BASED ON PCA

The feature extraction phase in any system is used for extracting suitable features from a dataset [19]. Sometimes it is necessary to use a PCA technique for extracting features to reduce the dimensionality of the dataset.

PCA is a statistical procedure that helps to identify the principal directions in which the data are modified. For example, in Fig. 2(a), assume the axes U and V represent a two-variable dataset that is measured in the X-Y coordinate system [23]. The main direction in which the data change is the U-axis, and the second important direction, which is orthogonal to the U-axis, is the V-axis, as shown in Fig. 2(b).

For each variable, if the (X,Y) coordinate is transformed into its corresponding (U,V) value, then the data are decorrelated, which means that the covariance between the U and V variables is zero. In other words, PCA can find the axis coordinate system defined by the main directions of the variance [24].

The parameter k in PCA represents the number of dimensions that is needed to reduce the features. There are two important factors related to the selection of k: the first factor is the average square projection, shown in Eq. (1). The second factor is the total variation in the data, shown in Eq. (2).

$$\sum\nolimits_{i=1}^{m} \left\| x^{(i)} - x_{approx}^{(i)} \right\|^2 \tag{1}$$

$$\sum\nolimits_{i=1}^{m} \left\| x^{(i)} \right\|^2 \tag{2}$$

Typically, k is chosen as the sum of the smallest values that satisfy the condition of Eq. (3). The k value selected (the number of eigenvalues) gives the difference between the original features and the reduced features, divided by the whole feature space; it should be less than or equal to

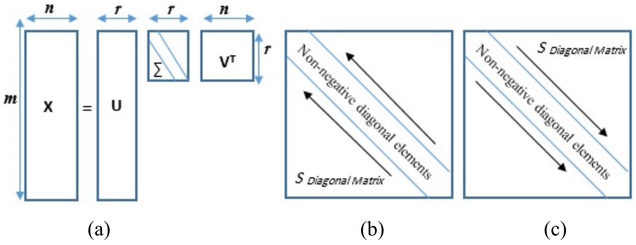
n
r
r
n
m
X
=
U
Σ
V^T
r
Non-negative diagonal elements
S Diagonal Matrix
Non-negative diagonal elements
S Diagonal Matrix
(a)
(b)
(c)

FIGURE 3. Singular-value decomposition. (a): standard, based on the SVD algorithms; (b) and (c): upward and downward evaluation of F Form.

the α value 0.01 [24].

$$\frac{\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^{2}}{\sum_{i=1}^{m}\left\|x^{(i)}\right\|^{2}}\leq 0.01 \tag{3}$$

In other words, an $\alpha \leq 0.01$ means that 99% of the variance can be recovered. It is possible to choose $\alpha \leq 0.05$ to retain 95% of the variance or $\alpha \leq 0.10$ for 90%. By using a MATLAB function, an S matrix with a diagonal of eigenvalues is found and by assuming $\alpha \leq 0.01$, then the sum of the k selected eigenvalues divided by the summation of all eigenvalues should be greater than or equal to 99%, as shown in Eq. (4) [25].

$$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} > 0.99 \tag{4}$$

C. DIMENSIONALITY REDUCTION BASED ON SVD

SVD is one of the most popular unsupervised data-mining algorithms and one of the most appropriate mapping tools used for mapping a high-dimensionality data space or vector space to other dimensions. Mathematically, let X be an m × n matrix and let the rank of X be r. The matrix rank is the largest number of rows (or columns) where no nonzero linear combination of rows is the zero vector (a set of such rows or columns is independent) [26]. Then, matrices U, Σ, and V are calculated as shown in Fig. 3.

SVD is a type of eigenvalue/eigenvector mechanics that uses a similar process of finding the singular value (eigenvector). Additionally, it is used to find the corresponding singular vectors (eigenvectors), which are mainly yielded by the matrix decomposition term. The terms 'singular vector' and 'eigenvector' will be used interchangeably, where the SVD of matrix A can be written as shown in Eq. (5) [27].

$$A = USV^{T} \tag{5}$$

where U is the orthogonal m×m matrix, and the columns of U are the eigenvectors of AA^T.

V is orthogonal to the n×n matrix, and the columns of V are the eigenvectors of the AA^T matrix. However, S is the set of diagonal eigenvalues (entities), which are also called the diagonal sigma values $\sigma_1,\ldots,\sigma_2$ and are computed based on the square roots of the nonzero eigenvalues of the AA^T and A^T A matrices. Both of them are the singular values of matrix A, and they occupy the first r places on the main diagonal of S, where r is defined as the rank of A. Based on Eq. (5), the connections with AA^T can be written as follows [28]:

TABLE 1. The selected dataset arrangement.

Class	Total	Training (80%)		Testing (20%)
		Training (80%)	Validation (20%)	
Healthy	134	85	22	27
High-grade	134	85	22	27
Low-grade	134	85	22	27
Total	402	255	66	81

$$AA^{T} = \left(USV^{T}\right)\left(VS^{T}U^{T}\right) = USS^{T}U^{T} \tag{6}$$

Similarly, A^TA can be written as follows:

$$A^{T}A = \left(VS^{T}U^{T}\right)\left(USV^{T}\right) = U^{T}USS^{T} \tag{7}$$

In Eq. (6), U must be the eigenvector matrix AA^T, and SS^T is the eigenvalue matrix that is placed in the middle, which is defined as the m×m matrix with eigenvalues $\lambda_1 = \sigma_1^2,\ldots,\lambda_r = \sigma_r^{2.}$.

However, based on the same method and using Eq. (7), U is defined as the eigenvector matrix for A^TA. The diagonal matrix S^T S has the same property $\lambda_1 = \sigma_1^2,\ldots,\lambda_r = \sigma_r^2$, which is an n×n matrix.

IV. METHODOLOGY

A. DATASET

MR brain images of low- and high-grade glioma tumours as well as the healthy brains of 130 patients, are used in this paper. There are 1467 axial-plane FLAIR MR Images in the dataset. All images are 256 x 256-pixel 8-bit RGB (colour images) in JPEG format. This open-access dataset is one of the most trusted datasets provided by TCIA (The Cancer Imaging Archive) [29], [30]. Only axial plane FLAIR MR images are used in this paper because the proposed system has to be able to achieve classification with the highest possible performance using only one type of MR image, as in [17], [31]–[33].

In the proposed system, three classes of MRI brain tumours are selected for classification. They are healthy, high-grade, and low-grade. Since the dataset has a different number of image samples in each class, which makes the dataset imbalanced, a random selection of an equal amount of data in each class is performed in this paper. Then, the selected images are divided into training, cross-validation, and testing datasets, as shown in Table 1. The subsets included patients with the various grades of tumours to ensure the robust performance of the classifier.

Five-fold cross validation is used. The best way to select the folds is sampling. The test size is set to a maximum of twenty percent of the dataset. This near equality allows for a more accurate evaluation of the resulting classifiers.

Then, the training set is divided into five folds; each fold has an equal number of randomly selected images. One fold is withheld for the validation step. These folds are used to train and validate the classifier, and the performance of the trained classifier is evaluated by the votes collected from each fold. Alternatively, MATLAB can automatically select the optimal scaling via a heuristic procedure using subsampling [34], [35].

B. THE FRAMEWORK OF THE PROPOSED SYSTEM

This paper proposes an intelligent system that includes six stages. A flowchart of the proposed system can be seen in Fig. 4. The proposed system is run using MATLAB version 15a on a 2.40-GHz Intel R core (TM) i7-4500U CPU with 16.0 GB of memory (RAM).

The following stages explain the proposed system of this paper in detail.

1) STAGE 1: PRE-PROCESSING

This stage includes steps to enhance the MR brain images and prepare them for the second stage, such as using the skull-removing technique and median and Gaussian filters to improve the speed and accuracy of diagnostics as well as tumour detection [36]–[38]; this is shown in Figs. 5 and 6.

2) STAGE 2: CLUSTERING BY LDI-MEANS

Image clustering is a technique used to separate an image into multiple slices and to find a region of interest. This paper used the LDI-Means clustering algorithm to segment the region of the tumour from the rest of the image, as explained in detail in a previous paper [38]. This method provided good results in terms of tumour segmentation and region of interest (ROI) detection for MR brain images. It is shown in Fig. 7.

3) STAGE 3: TUMOUR DETECTION AND LOCALIZATION

This stage includes finding the centre of the irregularity in terms of x and y values in addition to drawing a boundary box on the original image around the tumour, as shown in Fig. 8 [38].

4) STAGE 4: FEATURE EXTRACTION

In general, feature extraction is defined as a process used to reduce the amount of data required to describe a large set of data accurately for facilitating decision-making, such as pattern classification [34].

In many real-time applications, a mathematical equation is used to differentiate between two images. Two images may look the same to human eyes, but mathematically, each one will give a different result [39].

In this paper, 64 features are extracted from the tumour regions of MR brain images. Then, these features are used as inputs to the classifiers, which assign them to the class to which they belong. Accordingly, two types of features are extracted, which yield the structure of greyscale, symmetrical, and texture information for the segmented tumour. The two types of features are explained as follows:

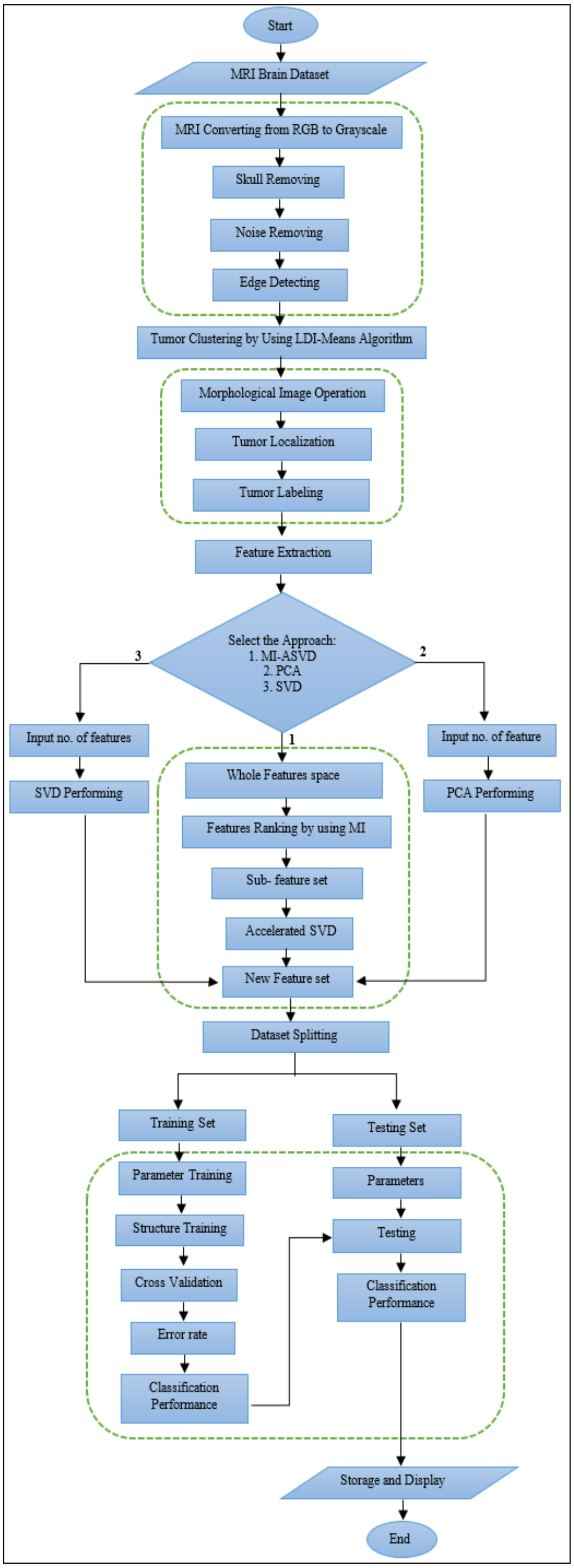

FIGURE 4. A flowchart of the proposed system in this paper.

a: GREYSCALE FEATURES

Five features are used in this paper: mean, variance, standard deviation, skewness, and kurtosis; they are defined

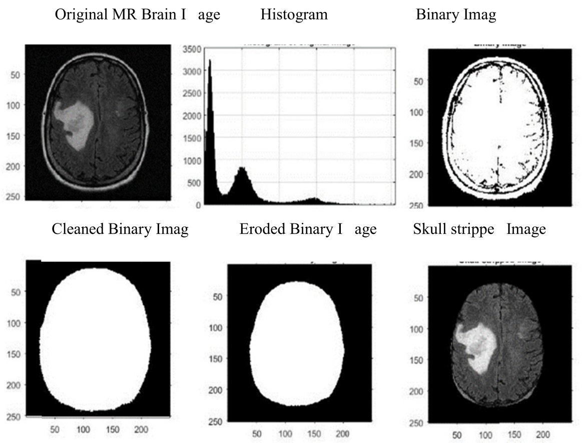

FIGURE 5. The steps of removing non-brain tissues in one image from the dataset as an example [38].

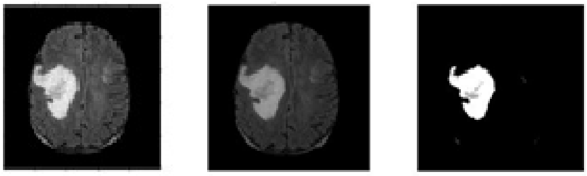

FIGURE 6. Smoothing and adjustment steps [38].

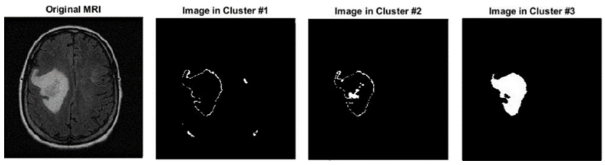

FIGURE 7. Clustering (segmentation), where the number of clusters K = 3 [38].

by Eqs. (8)-(12) [40]–[42].

$$Mean = \frac{\sum x}{n} \tag{8}$$

$$variance = \frac{\sum (x - mean)^2}{n-1} \tag{9}$$

$$Standard\ Deviation = \sqrt{\frac{\sum (x - mean)^2}{n-1}} \tag{10}$$

$$Skewness = \left(\frac{1}{variance}\right) \sum\nolimits_{x=1}^{m} \sum\nolimits_{y=1}^{n} (f\,(x,y) - mean)^3 \tag{11}$$

$$Kurtosis = \left(\frac{1}{variance}\right) \sum\nolimits_{x=1}^{m} \sum\nolimits_{y=1}^{n} (f\,(x,y) - mean)^4 \tag{12}$$

b: TEXTURE FEATURES

Several Haralick texture descriptors are extracted from each co-occurrence matrix and are computed according to the statistical properties of the image derived from the GLCM. They are the five features described in Eqs. (13)-(17) [41], [42].

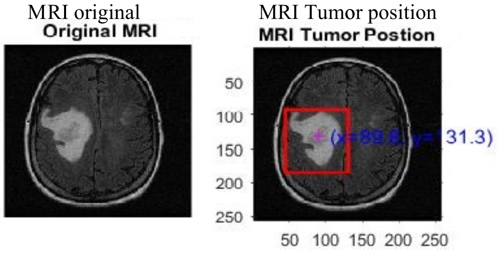

FIGURE 8. The position in terms of x and y of the brain tumour in one image from the dataset as an example [38]. (x = 89.6, y = 131.3).

$$Entropy = -\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{m} p\,(i,j)\,log(p(i,j) \tag{13}$$

$$Correlation = \frac{1}{(n-1)} \sum \frac{(x-\mu_x)(y-\mu_y)p(x,y)}{\sigma_x \sigma_y} \tag{14}$$

$$Contrast = \sum\nolimits_{x,y} |x-y|^2\, p(x,y)^2 \tag{15}$$

$$Energy = \sum\nolimits_{x,y} p(x,y)^2 \tag{16}$$

$$Homogeneity = \sum\nolimits_{x,y} \frac{p\,(x,y)}{1+|x-y|} \tag{17}$$

Additionally, the grey-level run-length matrix (GLRLM) feature extraction technique is used to extract 11 features depending on the derivation of the GLRLM for two-level high-frequency sub-bands of the discrete wavelet of the decomposed image. The first feature is for distance 1. The second is the degree, which is 0, 45, 90, and 135. The GLRLM feature extraction technique is used to isolate the relevant features from the tumour region, which leads to a good understanding of the MRI brain images. Then, the mean and variance of each feature are computed to extend the total number of features to 44 features [41].

5) STAGE 5: MI-ASVD

This stage is the most important part of this paper. It describes a novel method named MI-ASVD. This approach involves two techniques: feature selection based on MI theory and dimensionality reduction based on SVD. Additionally, this paper proposes a development of ordinary SVD that involves the selection of multiple eigenvalues, which has not been mentioned previously in the field of machine learning and data mining dimensionality reduction.

The whole idea of using MI-ASVD is to apply a feature selection method for re-ranking features and to then use the ordinary SVD technique as a non-biased thresholding algorithm; the accelerated SVD is named ASVD. This paper develops a mathematical MI model to evaluate the whole feature space. According to distance metric measurements, this model defines a threshold value of MI scoring that is based on the descent projection score of the majority feature score and the cumulative distribution function (CDF) value.

In general, MI is an information theory approach that describes and illustrates the relationship between two variables. Let us assume that S_i and S_j are two variables; the mutual information $I(S_i; S_j)$ measures how much information each variable has about the other [36]. The MI of two variables S_i and S_j, whose joint distribution is $H\left(S_i|S_j\right)$, is defined by Eq. (18) [43]:

$$I(S_i; S_j) = H\left(S_i\right) - H\left(S_i|S_j\right) = H\left(S_j\right) - H\left(S_j|S_i\right) \tag{18}$$

Basically, MI is defined as an absolute value that measures the information between two variables, as given in Eq. (19) [41]:

$$0 \le I(S_i; S_j) \le min(H\left(S_i\right) - H\left(S_i\right), H\left(S_j\right)) \tag{19}$$

MI is not a suitable measurement, since it is not based on a bounded range of the absolute value. However, it can allow the mutual information to measure the bounded range by normalizing the triangle inequality (distance metric) instead of the probability. Hence, let us assume that the mutual information-based triangle inequality metric determines the upper estimation of two variables by estimating the size of the sum of the two variables S_i and S_j, as shown in Eq. (20) [44].

$$\left|S_i - S_j\right| > \left|\left|S_i\right| - \left|S_j\right|\right| \tag{20}$$

By using the Euclidian distance space, the metric distance represents the norm of the inner product. Then, the triangle inequality using MI is defined as a given vector of S_i and S_j, where the inner product $\langle S_i, S_j\rangle$ is defined in Eq. (21) [43].

$$\begin{aligned} &\left\|S_i + S_j\right\|^2 \\ &= \langle S_i + S_j, S_i + S_j\rangle = \|S_i\|^2 \\ &\quad +\langle S_i, S_j\rangle + \langle S_i, S_j\rangle + \left\|S_j\right\|^2 \le \|S_i\|^2 + 2\langle S_i, S_j\rangle \\ &\quad + \left\|S_j\right\|^2 \le \|S_i\|^2 + 2\,\|S_i\|\,\left\|S_j\right\| + \left\|S_j\right\|^2 = (\|S_i\| + \|S_i\|) \end{aligned} \tag{21}$$

The last norm function is defined based on the Cauchy–Schwarz inequality. In this case, it becomes an equality for (dependent) linear variables if and only if the first variable S_i and the second S_j are nonnegative scalar numbers, as shown in Eq. (22) [43], [45]:

$$\langle S_i, S_j\rangle + \langle S_i, S_j\rangle \le 2\langle S_i, S_j\rangle \tag{22}$$

However, the MI is normalized by the possible maximum mutual score, as Eq. (23) shows [40]:

$$d_{CR}\left(S_i, S_j\right) = 1 - \frac{I\left(S_i, S_j\right)}{min(H\left(S_i\right), H\left(S_j\right))} \tag{23}$$

Then,

$$0 \le d_{CR}\left(S_i, S_j\right) \le 1 \tag{24}$$

Hence, the lower-bound range of the mutual scores shares the possible maximum score that is given by their entropies, as shown in Eq. (25) [45]:

$$d_{CR}\left(S_i, S_j\right) = 1 - \frac{I\left(S_i; S_j\right)}{max(H\left(S_i\right), H\left(S_j\right))} \tag{25}$$

It is based on the maximum entropy score for mutual normalization [44]–[46]. Thus,

$$1 - \frac{min\left(H\left(S_i\right), H\left(S_j\right)\right)}{max\left(H\left(S_i\right), H\left(S_j\right)\right)} \le d_{CR}\left(S_i, S_j\right) \le 1 \tag{26}$$

The MI-based distance metric returns zero if $S_i = S_j$, because it identifies and indicates features (variables) based on the maximum possible information gain, and the entropy is identical, as Eq. (27) shows:

$$H\left(S_i\right) = H\left(S_j\right) = I\left(S_i; S_i\right) \tag{27}$$

Otherwise, it depends on the MI score to indicate the selected variable such that:

$$\begin{cases} H\left(S_i\right) < H\left(S_j\right) \\ H\left(S_i\right) > H\left(S_j\right) \end{cases} \tag{28}$$

Hence, the distance between two variables $d\left(S_i, S_j\right)$ is based on the maximum entropy, which can be written as shown in Eq. (29) based on the complexity proposed by MacKay [44]:

$$d\left(S_i, S_j\right) = \frac{max\left(H\left(S_i\right), H\left(S_j\right)\right)}{max\left(H\left(S_i\right), H\left(S_j\right)\right)} \tag{29}$$

After applying the chain rules for the tested variable-based distance entropies $d\left(S_i, S_j\right)$, the MI-based distance metric can be written as Eq. (30) [45], [47].

$$H\left(S_i|S_j\right) \le H\left(S_k|S_j\right) + H\left(S_i|S_k\right) \tag{30}$$

where S_k results from applying the chain rules twice on the tested variables for three sources S_i, S_j.

The MI-based distance metric method used in this paper is defined in Algorithm (1).

Algorithm (1) shows how the MI is calculated by using two random variables and the Euclidean distance between each pair of features in the whole tested feature space and based on the normalized features using the maximum entropy score.

Since the MI-based distance metric has only positive scores, the threshold value is set to be 0.5 or less to achieve a smaller distance metric between the selected features, which is called the closed-interval feature score (semi-closed interval). The empirical distribution function is used to project the final MI-based distance score and find the final feature space. CDF is a cumulative distribution function of a real-valued random variable X as shown in Eq. (31):

$$F_X\left(x\right) = D(X \le x \in I_v \le 0.5) \tag{31}$$

where $P(X \le x)$ illustrates the distance of the whole feature space X, which selects only the values for which the MI score is less than or equal to x. The distance of X in the semi-closed interval is shown in Eq. (32):

$$Selection_{\,Threshold} = \forall_X F_X\left(a\right) F_X\left(b\right) = D\left(a < x \le b\right) \tag{32}$$

Algorithm 1 MI-Based Distance Metric Algorithm

```
Input: Space of all features S
Output: Mutual Scoring I
1. Repeat
2.    Calculate the mutual information for x and y
      $I(X,Y) = D(X,Y) = \frac{\max(H(\mathbf{x}),H(y))}{\max(H(y),H(\mathbf{x}))}$ where D is
      the Euclidean distance function
3.    Find $H(X|Y) \leq H(Z|Y) + H(X|Z)$
      where:
      X is the space of the first group of features, Y is the
      space of the second group of features and Z is the
      result of applying the chain rules twice on the tested
      variables with Eq. (30)
4. Call the next two variables
5. Until all the variables of the feature space have been
   used
6. Return vector I
7. End
```

Algorithm 2 ASVD Algorithm

```
Input: Data matrix X
Output: New Dimensions C
1. Repeat
2.    Construct the covariance matrix X from the
      decomposition according to:
         If $\frac{No.\ of\ Features}{No.\ of\ Samples} \geq 1$ then $Data \leftarrow X^T X$
                  else $Data \leftarrow XX^T$
         End if
3.    Calculate the d-dimensional mean vectors for each
      class from X.
        $XX^T = (USV^T)(USV^T)^T = (USV^T)(VSU^T)$
4.    Calculate V as an orthogonal matrix
           $(V^T V = I),\ XX^T = US^2U^T$
5.    Calculate the scatter matrices (between-class and
      within-class scatter X).
6.    Calculate $\sqrt[2]{\text{the eigenvalues of } XX^T} =$ singular
      values of X
7.    Compute the eigenvectors $(e_1, \ldots e_d)$ and the
      corresponding eigenvalues $(\lambda_{1,\ldots,}\lambda_d)$.
8.    Order the eigenvectors by decreasing eigenvalues.
9.    Select k eigenvectors with the smallest error (square
      root) from a $d \times k$ dimensional matrix W (where
      every column represents an eigenvector).
10.   Use this $d \times k$ eigenvector matrix to transform the
      samples into the new subspace. $Y \leftarrow X \times W$ where
      ($X$ is an $n \times d$ dimensional matrix, and $Y$ is the
      transformed $n \times k$ dimensional samples in the new
      subspace)
11. Until Convergence
12. End
```

In Eq. (32), the "less than or equal" sign illustrates the convention of the closest discrete distribution features falling between the lower bound distance score, which is 0 ("very close or the same"), and the upper bound score, which is 0.5 ("fairly close"). After performing MI scoring, an accelerated version of SVD is used. ASVD given in Algorithm (2).

6) STAGE 6: CLASSIFICATION BY A SIMPLIFIED RNN

The classification process aims to categorize data into individual classes. In a traditional ANN, each layer transfers the parameters to the next layer. More technically, each layer uses a feed-forward pass to feed the next layer directly until eventually many layers are reached. Typically, an ANN is a universal learning function that gradually increases the number of layers that are added to the structure. On the other hand, having a limited number of layers is still a major issue in the design of ANNs to improve accuracy. Hence, in some cases, the increase in the number of layers in an ANN results in a complex learning function that harms the ability of the universal learning function [48].

In contrast, deep learning is used to increase the number of layers with a simple learning function as a solution that increases the layer dimension in the ANN structure. The ANN therefore becomes more complex and deep than the original simple ANN. On the other hand, if the goal is to increase the number of layers but still use the simple ANN structure, it is possible to start at the point of eventual overfitting. This can show that a deep ANN is learned better than the regular ANN structure with an overfitting problem [46]. To overcome these issues, a residual neural network (RNN) is proposed as a modern ANN structure with the idea of a residual connection. Simply, an RNN is based on the connection between the previous layer and new layers. In this case, the simplified RNN tries to skip a connection from the previous layer to the next layer by avoiding the full connection and the complex learning function. This is one of the interesting solutions that retain the expanded neural network structure without an overfitting issue [48], [49]. The main diagram of the residual neural network (RNN) is illustrated in Fig. 9 [48].

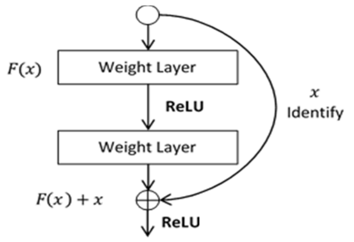

FIGURE 9. Single residual block diagram of a residual neural network [49].

From Fig. 8, it may be assumed that the difference between the input and the output (residual) is defined by Eq. (33) [49], [50]:

$$F(x) = Output - Input = H(x) - x \qquad (33)$$

TABLE 2. Architecture of the final classification approach.

Layer No.	Layer Type	Activation Function
Layer 2	Hidden Layer 1	Sigmoid
Layer 3	Hidden Layer 2	Sigmoid
Layer 4	Hidden Layer 3	Sigmoid
Layer 5	Classification Output	SoftMax
	Loss Function (optimization)	Cross Enropy

where the output is the new set of weights of the previous layer, and the input is the original set of weights of the next layer. By rearranging Eqs. (33) and (34) will be obtained [49], [50]:

$$H(x) = F(x) + x \tag{34}$$

In this case, the residual block is attempting to learn the true overall output $H(x)$. It can be seen from Fig. 8 above that RNN tries to learn the residual $F(x)$, since it has the identity connection (x) that comes from the same input x. In conclusion, the layers in the original ANN attempts to learn the output $H(x)$ only by learning and adjusting weights, while the residual neural network attempts to learn the true output $F(x)$ [49].

Conceptually, in the classification step, this paper used a simplified RNN methodology in a regular expanded neural network to reduce overfitting and achieve better accuracy. The input of the training formula consists of examples in the form of feature vectors with labels assigned to them. The goal of the classification algorithm is to learn to assign correct labels to new unseen samples of the general task [51].

The simplified RNN is used as a classifier based on the standard back-propagation algorithm in this paper, which contains one input layer, three hidden layers, and one output layer. The number of input nodes depends on the type of approach that is used. If a dimensionality reduction approach (PCA, SVD, or ASVD) is used, this means the input nodes depend on the total number of dimensions (k) that were selected before. The total number of output nodes depends on the total number of class labels, which is three in this paper.

Table 2 shows the simplified RNN architecture. In the all three hidden layers, every neuron has a sigmoid as an activation function. In the output layer, SoftMax is the main activation function. Standardly, the SoftMax activation function has a unique property that effectively indicates the output probability for each class in a multi-class classification problem, where the output is a value in [0-1]. SoftMax can choose the highest probability for each class, as shown in Eq. (35):

$$f(s)_i = \frac{1}{\sum_j^C e^{s_j}} \tag{35}$$

where s_j are the scores that are predicted by the network for each class in C.

The main optimization function that is used in the proposed simplified RNN design is the cross-entropy loss, as given in Eq. (36):

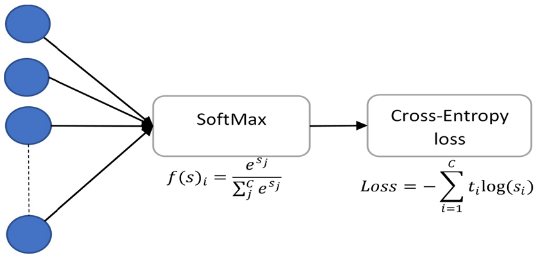
SoftMax
Cross-Entropy loss
$f(s)_i = \frac{e^{s_j}}{\sum_j^C e^{s_j}}$
$Loss = -\sum_{i=1}^{C} t_i \log(s_i)$

FIGURE 10. The architecture diagram for the classifier (SoftMax) and loss function (Cross-Entropy loss).

$$Loss = \sum\nolimits_{i=1}^{C} t_i log(s_i) \tag{36}$$

where t_i is the ground truth (label) of each class and s_i is the probability score (predicted) for each class.

Since the proposed network is designed for a multi-class classification problem, the SoftMax activation function with cross-entropy loss is used (as SoftMax-Entropy loss). The proposed network is trained to obtain the probability output over the class (c_j) for each tumour image in the training phase for all classes, as shown in Eq. (37):

$$f(s)_i = \frac{e^{s_j}}{\sum_j^C e^{s_j}} Loss = -\sum\nolimits_{i=1}^{C} t_i log(s_i) \tag{37}$$

For multi-class classification labelling, one-hot coding is used; in this case, only the positive classes c_p remain in the loss function (main term). One element in the loss function, that is, the target vector (ground truth), is as follows:

$$t_i = t_p \tag{38}$$

Based on the target labels, the elements for which the summation is zero are discarded. According to this assumption, the optimization loss function is written as follows:

$$Loss = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right) \tag{39}$$

The architecture diagram for both the classifier (SoftMax) and loss function (Cross-Entropy loss) is shown in Fig. 10.

The proposed network is trained by using the following parameters:

- The initial learning rate parameter is 0.0001.
- The momentum factor is used to adjust the step size for the global minimum coverage by setting it to 0.9.
- The learning patch size is 16.
- The epoch size is 20.
- The iteration number for each epoch is 500.
- The dataset is augmented by pre-processing using a Gaussian filter, as shown in Eq. (40) below:

$$K_\sigma(t) = exp\left(\frac{1}{2\sigma^2}\sum\nolimits_{i=1}^{n} t_i^2\right) \tag{40}$$

TABLE 3. Confusion matrix.

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

C. EVALUATING PARAMETERS

In general, a confusion matrix is a technique for summarizing the performance of a classification algorithm, as shown in Table 3 [47].

True Positive (TP) refers to the correct detection of positive cases, true negative (TN) refers to the correct detection of negative cases, false positive (FP) refers to the incorrect classification of positive cases in the negative class and false negative (FN) refers to the incorrect classification of negative cases in the positive class.

The evaluation parameters of the MRI classification system are calculated by using the following measures [47], [52], [53]:

$$Accuracy\ (Recognition\ Rate) = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{41}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \times 100 \tag{42}$$

$$Precision\ (Specificity) = \frac{TN}{TN+FP} \times 100 \tag{43}$$

$$F1 - Measurement = 2 \times \frac{2TP}{2TP+FP+FN} \tag{44}$$

$$FalseAlarm = \frac{FN}{FN+TN} \tag{45}$$

In this paper, a confusion matrix is 3 × 3.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Four sets of experiments are performed to test the performance and robustness of the proposed system. In all experiments, three classes are classified and the simplified RNN is used as a classifier.

- Experiment 1: using the whole feature space of 64 features, which are extracted from the segmented ROIs.
- Experiment 2: using the most significant features (13 features), which are obtained as a result of using MI-ASVD.
- Experiment 3: using the PCA technique to decrease the number of features from 64 to $k = 13$, 26 and 52.
- Experiment 4: using the SVD technique to decrease the number of features from 64 to $k = 13$, 26 and 52.

The experimental results are obtained based on three criteria for comparison. Firstly, with the all features (experiment 1), as shown in Tables 4 and 5. Secondly, with the PCA and SVD for different values of k (experiments 3 and 4), as shown in Tables 4 and 5. Thirdly, with some published studies, as shown in Table 6.

TABLE 4. The classification accuracy.

Approach	Training	Testing	No of Features
All features + simplified RNN	77.58	79.20	64
MI-ASVD + simplified RNN	**92.69**	**94.91**	**13**
PCA + simplified RNN	86.21	88.02	13
	75.22	76.17	26
	73.74	74.64	52
SVD + simplified RNN	85.89	87.71	13
	72.90	73.85	26
	72.40	73.31	52

TABLE 5. The classification performance results.

Approach	Criterion	Training	Testing
All Features + simplified	Recall	76.62	77.70
	Precision	77.12	78.17
	F1-Measure.	70.56	71.10
	Detection Rate	71.79	72.75
	False Alarm	25.69	23.98
MI-ASVD + simplified RNN	**Recall**	**96.69**	**96.89**
	Precision	**96.70**	**96.37**
	F1-Measure.	**96.49**	**96.80**
	Detection Rate	**89.84**	**91.36**
	False Alarm	**7.08**	**5.71**
PCA + simplified RNN	Recall	86.45	86.53
	Precision	86.46	86.54
	F1-Measure.	86.45	86.63
	Detection Rate	83.56	85.33
	False Alarm	13.37	11.61
SVD + simplified RNN	Recall	86.44	86.42
	Precision	86.34	86.62
	F1-Measure.	86.24	86.53
	Detection Rate	83.25	85.02
	False Alarm	13.68	11.92

Table 4 shows the classification accuracy of experiments 1, 2, 3 and 4. The second row in Table 4 shows the accuracy of the proposed algorithm MI-ASVD, which achieved 92.69% on the training set and 94.91% on the testing set. According to the MI-ASVD approach, 13 features were selected automatically from the total number of features, which is 64. The names and feature ranks are shown in Table 7.

Based on the 13 features obtained, this paper manually selected K = 13 to be the required reduction for PCA and SVD. PCA achieved an accuracy of 86.21% on the training set and 88.89% on the testing set, whereas SVD achieved an accuracy of 85.89% on the training set and 87.71% on the testing set.

Moreover, in terms of measuring the performance of PCA and SVD, different numbers of features were used. Twenty-six (double of 13) and 52 (double of 26) features were used. When using 26 features, the accuracy values are 75.22% and 76.17% on training and testing, respectively, for PCA; whereas, they are 72.90% and 73.85% on training and testing, respectively, for SVD, as shown in Table 3. In the case of using 52 features, the accuracy values are 73.72% and 74.64% on training and testing, respectively, for PCA; whereas, they

TABLE 6. Brain tumor classification by different feature selection and machine learning algorithms for comparison with the proposed system.

Study	Modality / Database	Method	No. of classes	Sensitivity %	Specificity %	Accuracy %
Zacharaki *et al.* [16]	Multiple modalities / Private dataset	SVM-RFE + SVM	2 classes (LLG and HGG)	84.6	95.5	87.8
Jones *et al.* [15]	Multiple modalities / Private dataset	SVM + D-SEG spectra	5 classes (LGG, GBM, cGBM, MET and MEN)	> 90	> 97	94.7
Gupta *et al.* [32]	Flair / Fortis Memorial Research Institute	DWT + PCA + CART, Random Forest, KNN Linear SVM	2 classes (Normal and Abnormal)	80, 80, 80, 80	88, 96, 80, 96	84, 88, 80, 88
Sachdeva *et al.* [11]	T1 / PGIMER	CBAC + PCA + ANN	6 classes (AS, GBM, MED, MEN, MET and NR)	N/A	N/A	91
Hsieh *et al.* [54]	T1 / TCIA	Logistic Regression	2 classes (LLG and GBM)	82	90	88
Yang et al. [55]	Multiple modalities / Private dataset	D-SEG + SVM	2 classes (GBM and MET)	94.4	90	91.6
Soltaninejad *et al.* [17]	Flair / MICCAI BRATS 2012	mRMR + SVM, ERT	3 classes (grade II, grade III and grade IV)	82.7, 88.0	83.7, 89.0	0.83, 0.88 (Dice)
Proposed system	**Flair / TCIA**	**MI-ASVD + simplified RNN**	**3 classes (Healthy, LLG and HGG)**	**96.8**	**96.3**	**94.9**

Abbreviations: cGBM, cystic GBM; LGG, low-grade glioma; MEN, meningioma; MET, metastasis; HGG, high grade glioma; PGIMER, Department of Radiodiagnosis, Postgraduate Institute of Medical Education & Research, India

TABLE 7. Names and feature ranks based on the MI-ASVD approach.

No.	Original No.	Feature Name	Ranking Score (MI)
1	18	standard deviation of the correlation	0.155009
2	16	standard deviation of the contrast	0.155009
3	20	standard deviation of the dissimilarity	0.152422
4	27	mean of the entropy	0.150504
5	52	standard deviation of the inverse difference normalized	0.146310
6	29	mean of the homogeneity	0.144260
7	45	mean of the difference entropy	0.143850
8	41	mean of the sum entropy	0.140517
9	51	mean of the inverse difference normalized	0.139570
10	39	standard deviation of the sum variance	0.137944
11	30	standard deviation of the homogeneity	0.134419
12	19	mean of the dissimilarity	0.132471
13	42	standard deviation of the sum entropy	0.130517

are 72.40% and 73.31% on training and testing, respectively, for SVD.

Table 6 shows a comparison with other studies in this field. According to the comparison, it is found that the performance and accuracy of the proposed system is suitable for providing a meaningful estimation of brain tumour classification in real time and provides good precision to detect this class of brain tumours.

The average quality measurement results, such as recall, precision, F-measurement, detection rate, and false alarm of the proposed system during the training and testing phases, are shown in Table 5.

Additionally, it can be seen in Fig. 11 that the MI-ASVD approach (blue loss function line) achieves the lowest score during the training phase within 500 epochs compared with the other dimensionality reduction approaches PCA and SVD, in grey and orange, respectively, as well as in comparison with the whole feature space, in purple.

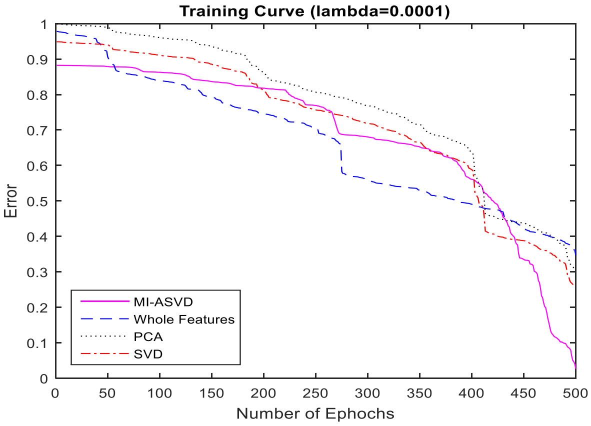

FIGURE 11. Loss function scores of the classification training phase by using all features, MI-ASVD, PCA and SVD.

VI. CONCLUSION

The main contribution of this paper is to design an automated system that can detect and classify grades of brain gliomas. The classification process in the proposed system depends only on the most significant features obtained by using MI-ASVD. It is a new method combining the mutual information with Singular Value Decomposition. It automatically selects the significant subset of features and provides a good distribution in finding the best features to use as an input to the main classifier.

Here, a hybrid methodology, MI-ASVD with a simplified RNN has been used in addition to LDI-Means clustering [38] proposed in a previous paper by the same authors. This method presents both an unsupervised and also supervised learning approaches that can make a useful CAD system to help speed up the diagnostic procedure and decrease diagnostic errors. This combination gives an accurate result for identifying brain tumours by achieving 94.91% accuracy on the testing dataset, whereas the highest accuracy using PCA was 88.02% and that using SVD was 87.71%. The experimental results of the proposed system demonstrates the effectiveness of using MI-ASVD technique which identified the robust features to recognize the class of the tumour excellently and to save time.

Regarding the selected features, there are some limitations that the proposed system may face. MI-ASVD works perfectly on the dataset used in this paper, but there are a variety of other datasets, which may present different difficulties and challenges, and the proposed model could be unsuitable for them. Hence, this paper proposes using different classifiers to increase the accuracy of associating different segmentation and feature extraction methods with other clinical cases by using a large dataset to cover different scenarios and overcome these limitations. Additionally, possible future work includes:

- Using T1-weighted and T2-weighted MR images, because this paper used only FLAIR-weighted MR images.
- Using 3-D VOIs for evaluation, which could be more convincing.
- Increasing the number of classes, which could provide more information on the grades of glioma tumours.
- Using deep neural networks, because despite some successes, deep neural network applications remain relatively unexplored in the neuroimaging field [56].

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable suggestions and comments to improve this manuscript.

REFERENCES

[1] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[2] K. T. Keerthana and S. Xavier, "An intelligent system for early assessment and classification of brain tumor," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Coimbatore, India, Apr. 2018, pp. 1265–1268.

[3] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018, doi: 10.1109/TMI.2017.2785879.

[4] V. Zeljkovic, C. Druzgalski, Y. Zhang, Z. Zhu, Z. Xu, D. Zhang, and P. Mayorga, "Automatic brain tumor detection and segmentation in MR images," in *Proc. Pan Amer. Health Care Exchanges (PAHCE)*, Brasilia, Brazil, Apr. 2014, p. 1.

[5] M. Soltaninejad, X. Ye, G. Yang, N. Allinson, and T. Lambrou, "Brain tumour grading in different MRI protocols using SVM on statistical features," in *Proc. 18th Conf. Med. Image Understand. Anal. (MIUA)*, Egham, U.K., Jul. 2014, pp. 259–264.

[6] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowl.-Based Syst.*, vol. 140, pp. 103–119, Jan. 2018, doi: 10.1016/j.knosys.2017.10.028.

[7] E. Tsolaki, "Clinical decision support systems for brain tumor characterization using advanced magnetic resonance imaging techniques," *World J. Radiol.*, vol. 6, no. 4, pp. 72–81, Apr. 2014, doi: 10.4329/wjr.v6.i4.72.

[8] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994, doi: 10.1109/72.298224.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[10] V. Kumar, J. Sachdeva, I. Gupta, N. Khandelwal, and C. K. Ahuja, "Classification of brain tumors using PCA-ANN," in *Proc. World Congr. Inf. Commun. Technol.*, Mumbai, India, Dec. 2011, pp. 1079–1083.

[11] J. Sachdeva, V. Kumar, I. Gupta, N. Khandelwal, and C. K. Ahuja, "Segmentation, feature extraction, and multiclass brain tumor classification," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1141–1150, Dec. 2013, doi: 10.1007/s10278-013-9600-0.

[12] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," *IEEE Trans. Med. Imag.*, vol. 27, no. 5, pp. 629–640, May 2008, doi: 10.1109/TMI.2007.912817.

[13] L. Fang, H. Zhao, P. Wang, M. Yu, J. Yan, W. Cheng, and P. Chen, "Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data," *Biomed. Signal Process. Control*, vol. 21, pp. 82–89, Aug. 2015, doi: 10.1016/j.bspc.2015.05.011.

[14] N. Hoque, H. A. Ahmed, D. K. Bhattacharyya, and J. K. Kalita, "A fuzzy mutual information-based feature selection method for classification," *Fuzzy Inf. Eng.*, vol. 8, no. 3, pp. 355–384, Sep. 2016, doi: 10.1016/j.fiae.2016.09.004.

[15] T. L. Jones, T. J. Byrnes, G. Yang, F. A. Howe, B. A. Bell, and T. R. Barrick, "Brain tumor classification using the diffusion tensor image segmentation (D-SEG) technique," *Neuro-Oncol.*, vol. 17, no. 3, pp. 466–476, Mar. 2014, doi: 10.1093/neuonc/nou159.

[16] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1609–1618, Dec. 2009, doi: 10.1002/mrm.22147.

[17] M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe, and X. Ye, "Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 183–203, Feb. 2017, doi: 10.1007/s11548-016-1483-3.

[18] M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe, and X. Ye, "Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels," *Comput. Methods Programs Biomed.*, vol. 157, pp. 69–84, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.003.

[19] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, Edinburgh, Scotland, Jul. 2017, pp. 506–517.

[20] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*. [Online]. Available: http://arxiv.org/abs/1811.02629

[21] P. Mahindrakar and M. Dr Hanumanthappa, "Data mining in healthcare: A survey of techniques and algorithms with its limitations and challenges," *Int. J. Eng. Res. Appl.*, vol. 3, no. 6, pp. 937–941, Nov.-Dec. 2013.

[22] R. M. Gray, *Entropy and Information Theory*. New York, NY, USA: Springer-Verlag, 1990.

[23] G. Zeng, "A unified definition of mutual information with applications in machine learning," *Math. Problems Eng.*, vol. 2015, pp. 201–874, Mar. 2015, doi: 10.1155/2015/201874.

[24] L. I. Smith. *A Tutorial on Principal Component Analysis*. Accessed: Feb. 26, 2002. [Online]. Available: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[25] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Zolghadri Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognit.*, vol. 44, no. 7, pp. 1357–1371, Jul. 2011, doi: 10.1016/j.patcog.2010.12.015.

[26] W. A. Shehab and Z. Al-qudah, "Singular value decomposition: Principles and applications in multiple input multiple output communication system," *Int. J. Comput. Netw. Commun.*, vol. 9, no. 1, pp. 13–21, Jan. 2017, doi: 10.5121/ijcnc.2017.9102.

[27] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[28] A. Smola and S. V. N. Vishwanathan, *Introduction to Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[29] L. Scarpace, A. E. Flanders, R. Jain, T. Mikkelsen, and D. W. Andrews. (2015). *Data From REMBRANDT*. [Online]. Available: http://dx.doi.org/10.7937/K9/TCIA.2015.588OZUZB

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[31] G. D. Shah, S. Kesari, R. Xu, T. T. Batchelor, A. M. O'Neill, F. H. Hochberg, B. Levy, J. Bradshaw, and P. Y. Wen, "Comparison of linear and volumetric criteria in assessing tumor response in adult high-grade gliomas1," *Neuro-Oncology*, vol. 8, no. 1, pp. 38–46, Jan. 2006, doi: 10.1215/S1522851705000529.

[32] T. Gupta, T. K. Gandhi, R. K. Gupta, and B. K. Panigrahi, "Classification of patients with tumor using MR FLAIR images," *Pattern Recognit. Lett.*, Oct. 2017, doi: 10.1016/j.patrec.2017.10.037.

[33] F. Raschke, T. R. Barrick, T. L. Jones, G. Yang, X. Ye, and F. A. Howe, "Tissue-type mapping of gliomas," *NeuroImage: Clin.*, vol. 21, Jan. 2019, Art. no. 101648, doi: 10.1016/j.nicl.2018.101648.

[34] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, Mar. 2010, doi: 10.1109/TPAMI.2009.187.

[35] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40–79, Oct. 2018.

[36] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[37] M. K. Chung. *Gaussian Kernel Smoothing, Lecture Notes*. Accessed: Sep. 2, 2011. [Online]. Available: http://contents.kocw.or.kr/document/wcu/2012/Seoul/ChungMooK/02.pdf

[38] Z. A. Al-Saffar and T. Yildirm, "An optimized clustering approach for tumor segmentation using local difference of intensity level in MR brain images," in *Proc. Innov. Intell. Syst. Appl. (INISTA)*, Thessaloniki, Greece, Jul. 2018, pp. 1–8.

[39] S. Kuanar, V. Athitsos, D. Mahapatra, K. R. Rao, Z. Akhtar, and D. Dasgupta, "Low dose abdominal CT image reconstruction: An unsupervised learning based approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 1351–1355.

[40] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM," *Int. J. Biomed. Imag.*, vol. 2017, Mar. 2017, Art. no. 9749108, doi: 10.1155/2017/9749108.

[41] A. Attig and P. Perner, "A comparison between Haralicḱs texture descriptor and the texture descriptor based on random sets for biological images," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Berlin, Germany: Springer, 2011, pp. 524–538.

[42] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[43] S. Kullback, *Information Theory and Statistics*. Gloucester, MA, USA: Peter Smith, 1978.

[44] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[45] A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. E. S. Socolar, "Mutual information in random Boolean models of regulatory networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 77, no. 1, Jan. 2008, Art. no. 011901, doi: 10.1103/PhysRevE.77.011901.

[46] D. Keys, S. Kholikov, and A. A. Pevtsov, "Application of mutual information methods in time–distance helioseismology," *Sol. Phys.*, vol. 290, no. 3, pp. 659–671, Mar. 2015, doi: 10.1007/s11207-015-0650-y.

[47] P. Shanthakumar and P. Ganeshkumar, "Performance analysis of classifier for brain tumor detection and diagnosis," *Comput. Electr. Eng.*, vol. 45, pp. 302–311, Jul. 2015, doi: 10.1016/j.compeleceng.2015.05.011.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[49] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, "The reversible residual network: Backpropagation without storing activations," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, Dec. 2017, pp. 2214–2224.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[51] V. Gupta and K. S. Sagale, "Implementation of classification system for brain cancer using backpropagation network and MRI," in *Proc. Nirma Univ. Int. Conf. Eng. (NUiCONE)*, Ahmedabad, India, Dec. 2012, pp. 1–4.

[52] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," in *Proc. North East SAS Users Group Conf.*, Baltimore, MD, USA, Nov. 2010, pp. 1–10.

[53] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, Feb. 2008.

[54] K. L.-C. Hsieh, C.-M. Lo, and C.-J. Hsiao, "Computer-aided grading of gliomas based on local and global MRI features," *Comput. Methods Programs Biomed.*, vol. 139, pp. 31–38, Feb. 2017.

[55] G. Yang, T. L. Jones, F. A. Howe, and T. R. Barrick, "Morphometric model for discrimination between glioblastoma multiforme and solitary metastasis using three-dimensional shape analysis," *Magn. Reson. Med.*, vol. 75, no. 6, pp. 2505–2516, Jun. 2016, doi: 10.1002/mrm.25845.

[56] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive analysis of working memory load from eeg, by a deep recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2576–2580, doi: 10.1109/ICASSP.2018.8462243.

ZAHRAA A. AL-SAFFAR received the B.Sc. degree in biomedical engineering from the University of Baghdad, Baghdad, Iraq, in 2003, and the M.Sc. degree in medical engineering from Al-Nahrain University, Baghdad, in 2010. She is currently pursuing the Ph.D. degree with the Department of Electronics and Communications Engineering, Yildiz Technical University, Istanbul, Turkey. From 2010 to 2014, she worked as an Assistant Lecturer with the Department of Biomedical Engineering, Al-Khwarizmi College of Engineering, University of Baghdad. Her research interests include biomedical engineering, machine learning, intelligent systems, and medical image processing.

TÜLAY YILDIRIM (Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics and communication engineering from Yildiz Technical University, Istanbul, Turkey, in 1990 and 1992, respectively, and the Ph.D. degree in electrical and electronics engineering from the University of Liverpool, U.K., in 1997. She is currently a Full Professor with the Department of Electronics and Communications Engineering, Yildiz Technical University. She has authored over 200 publications in journals, conferences, and books. Her research interests include artificial intelligence, intelligent systems, cyber-physical systems, biomedical instrumentation, biometric person identification and verification systems, electronic circuits and systems, and artificial neural networks.