SUPERVISED CLASSIFICATION OF NORMAL AND TUMOROUS BRAIN MR IMAGES USING MACHINE LEARNING SCHEMES

Lim Jia Qi*, Norma Alias, Farhana Johar
Department of Mathematical Sciences, Faculty of Science, 81310 UTM Johor Bahru, Johor, Malaysia
*Corresponding author
Email address: ykj10@hotmail.com (Lim Jia Qi), norma@ibnusina.utm.my (Norma Alias), farhanajohar@utm.my (Farhana Johar)

Manual interpretation of these huge amounts of image volumes are susceptible to inter-reader variability and human error. Thus, accurate automated CAD scheme is highly desirable in clinical pathological diagnosis. In this research, plethora of machine learning paradigms (e.g. feature extraction, dimensionality reduction and supervised classification methods) were explored, evaluated, compared and analyzed to identify the optimal pathway for brain MR images (normal vs neoplastic) binary classification task. External validation dataset was used to test the generalizability of the optimal predictive models implemented. Relevant and informative features were selected to construct cross-validated decision tree and eventually simple rule set was built based on the decision tree. The experimental results show that almost all pattern recognition paradigms achieve high accuracy with careful selection of number of attributes. LDA+ELM with 55 features are the optimal pipelines which achieve perfect classification when training and test data are of same source; and achieving (accuracy=97.5%, AUC=0.989, sensitivity=95% and specificity=100%) under balanced test dataset; (accuracy=99.5%, AUC=0.988, sensitivity=95% and specificity=100%). Cross-validated decision tree model also shows comparable performance: accuracy=98.8%, AUC=99.1%, sensitivity=99.6% and specificity=98.2%. Three highly relevant and robust attributes are visualized and selected for construction of decision tree models and finally a rule sets are read directly off the decision tree. This rule sets can potentially serve as fast and accurate classification algorithm.

Keywords: Computer aided diagnosis (CAD), brain magnetic resonance imaging (MRI) scans, feature extraction, feature reduction, classifiers, classification rule.

1. Background

Brain tumor is a potentially life threatening medical condition and can inflict people, regardless of their age, ethnicity or gender. Early detection and diagnosis can lead to better understanding and accurate characterization of brain tumor, which is of paramount importance for pre-treatment planning and surgery and ultimately increase the chances of survival of patients.

With the rapid advancement of biomedical image analysis technologies, coupled with development of statistical modeling algorithms, biomedical imaging has emerged as a powerful tool in diagnosis procedure aside from gold standard biopsy (resection of tissue samples) nowadays. Among the available imaging protocols, magnetic resonance imaging (MRI) is one of the most promising one due to its several properties:1) no harmful radiation, 2) multi-spectral high resolution images, 3) superior soft tissue differentiation, and 4) high contrast. However, manual interpretation of this enormous volumes of images can be tedious, time consuming and prone to both observers' fatigue and variability in opinions among human experts. Therefore, automated or at least interactive computer assisted diagnosis (CAD) system is of great demand to facilitate accurate and quick detection and diagnosis of brain tumor [1].

By utilizing and integration of high quality imaging tool, like MRI and machine learning approaches, classification of brain MR images (normal vs neoplastic) using either supervised or unsupervised learning methods has been proposed in substantial number of papers [2, 3]. Normally,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

supervised method achieves higher accuracy, but unsupervised method is also desirable in biomedical image analysis since annotated images is not readily available [4]. Despite presence of vast amount of advanced pattern recognition algorithms, brain MR images classification still remain an open challenge as there is no universally accepted optimum radiomics and classifiers. This is primarily due to differences in brain anatomical structures, wide variety of MRI parameters, imaging artifacts, imaging features that create ambiguity [5] and so on. Furthermore, some methods proposed in the literature are subject to overfitting (a condition where classifiers fit well to training data, but fail to generalize to unseen data) and there is limited clinically accepted validation process to support the superior classification accuracy obtained [6].

In view of various available machine learning algorithms being proposed and no widely accepted best method, the authors intend to conduct comparative study of several feature extraction, feature reduction and classifiers in brain MR binary classification problem (normal vs tumorous). The performance of these methods will be analyzed by using six performance measures: test accuracy, sensitivity, specificity, area under the curve of receiver operating curve (AUC), training time and test time. This hierarchical procedures can be named as radiomics [7]. According to [8], it is impossible to compare and validate performance of learning algorithms without the measure of variance. Thus, training and test phases for each experiments are conducted 30 times to produce unbiased performance measures. In addition, in order to gauge and analyze the generalizability of the machine learning paradigm, additional images from completely different sources are included as test dataset.

The contributions of this research are five-fold: 1) large amount of samples (thousands of brain MR images) are used as input data for feature engineering as well as classifiers training and evaluation regarding classification of pathologic brain MR images, 2) perform repeated sampling of training and test dataset for reliable and unbiased evaluation, analysis and comparison of each supervised learning methods, 3) extra independent test dataset (external validation set) is utilized to gauge and validate the generalizability of the optimal machine learning pipeline aforementioned , 4) determination of relevant and informative features that can provide new insight and perspective for this binary classification problem, 5) construction of compact classification rule, which can be readily used in CAD system.

2. Methods

The overall pipeline of the research framework is outlined in Fig. 1. This experiments were performed using MATLAB R2017a on Intel® Core™ i5 processor with 3.89GB usable RAM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

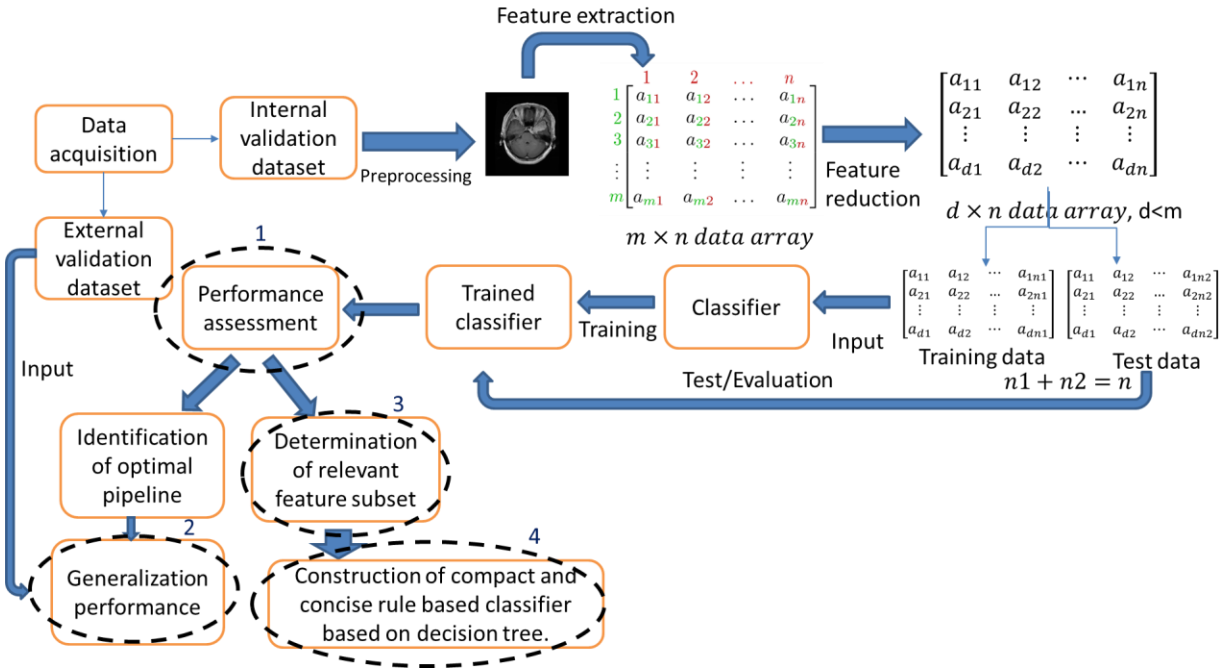


Fig. 1. Schematic diagram of research framework (brain tumor classification). The dashed circles refer to the important findings of this research.

2.1. *Data acquisition*

The brain MR images were all downloaded from some popular publicly available online database. The justification of using multiple sources of tumorous and normal brain T1-weighted MR images is to prevent the feature extracted and thus the information learned by the classifiers to be affected by or related to the way the MR images are acquired, like the types and parameter settings of various MRI machine. In other words, this step can be crucial to find distinctive, repeatable and robust feature or features subset in differentiating between tumorous and normal MR images regardless of the sources of data. The dataset used are downloaded from some online repositories and are listed as below:

1) A total of 3064 slices of T1-weighted contrast enhanced MR images from 233 patients was downloaded from https://figshare.com/articles/brain_tumor_dataset/1512427. There are three kinds of brain tumor in the MR images downloaded, namely meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices). The brain T1-weighed CE-MRI dataset was acquired from Nanfang Hospital, Guangzhou, China, and General Hospital, Tianjing Medical University, China, from year 2005 to 2010. The images have an in-plane resolution of 512×512 with pixel size 0.49×0.49 mm^2. The slice thickness is 6 mm and the slice gap is 1 mm [9, 10]. **1500** slices of these images were utilized as tumorous images.
2) Another set of tumorous brain MR image volume were acquired from [11, 12]. A total of **775** slices of image containing tumor were extracted.
3) Tumorous brain image volumes [13] were downloaded from http://nist.mni.mcgill.ca/?page_id=672. Pre- and post-operative T1-weighted MR with 63

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

gadolinium have been acquired from 14 brain tumor patients at the Montreal Neurological Institute in 2010. A total of 100 MR images were extracted.

4) 581 brain T1 weighted MR image volumes from normal subjects was downloaded from http://brain-development.org/ixi-dataset/. A total of **1500** 2D MR images were extracted.
5) **900** 2D normal brain images were extracted from Neurofeedback skull-stripped (NFBS) repository (http://preprocessed-connectomes-project.org/NFB_skullstripped/) [14].
6) 18 T1-weighted brain image volumes [15] were downloaded from https://www.nitrc.org/frs/?group_id=48. 900 2D MR images are extracted.

Dataset 1-3 consists of tumorous images, whereas dataset 4-6 was made up of normal images. Dataset 1, 2, 4, 5 will be used for training and test classifiers, while dataset 3 and 6 will serve as *external validation dataset* to evaluate the generalization performance of optimal machine learning schemes [16].

2.2. *Image preprocessing*

Image preprocessing on every dataset is performed to increase signal to noise ratio and enhance the visual quality of MR images. Steps of preprocessing of brain MR images is presented in Fig. 2.

Fig. 2. Process flow of brain MR images preprocessing.

In this research, first stage of MR images preprocessing is bias field estimation using second order parametric surface fitting method by Levenberg-Marquadt algorithm, in reference to work of [17]. The steps of the algorithm are:

1. Different from previous work of [17], which suggested the selection of data points from background image, data points with *normalized pixel intensity higher than pre-specified threshold 0.3* are selected. Their coordinates and gray level values are stored in data matrix $D = (xi, yi, G)$, $i = 1,2, \dots, Ng$, where Ng is the number of gray level.
2) First order parametric equation is selected for the fitted surface. Low order polynomial results in smooth surfaces and easy estimation of parameters. Non-linear least squares method: Levenberg-Marquadt algorithm is utilized to estimate the coefficients.
3) Use the fitted equation to generate final bias field signal by linear projection of original image with the coefficients estimated in step 2.
4) Perform element-wise division of original image with the generated bias field image in step 3 to yield the bias field removed MR images.

Next, 5 × 5 median filter is applied to retain important edges and eliminate noise. Then, histogram stretching or also known as inner cropping is performed on the image. The goal is to increase contrast without modifying shape of the original histogram (distribution of gray level).

2.3. Feature extraction

Next, the skull stripped images will undergo feature extraction, a process in which input data (MR

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

images pixel intensity) are transformed into much lower dimensional feature subspace (feature vectors). Ideally, these features should be informative, relevant and non-redundant. The main goal of feature extraction is to reduce computational cost (increase convergence speed during training) and data storage requirements without compromising the accuracy of classification models. In this research, 6 types of feature extraction algorithms are employed: first order statistical feature, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), histogram of oriented gradients (HOG), local binary pattern (LBP), and discrete wavelet packet transform (DWPT). It should be noted that the images come with different dimensions, thus before feature extraction (except first order statistical features), all the images were resized to 256×256. The number of features extracted from each method are summarized in Table 1. After the process of feature extraction, it was found that there are 161 LBP features of that contain all zero for every samples. In view of this, these non-informative features are removed. So, the final whole data dimension is 4675×2527.

Table 1. Feature extraction and its number of features.

Feature extraction methods	Number of features
First order statistical features	6
GLCM	36
GLRLM	28
HOG	1764
LBP	105
DWPT	588

2.3.1 First order statistical features

Textural features that will be computed include mean, variance, skewness, kurtosis, energy, range and entropy. Let G be the gray level of an image. The relative frequency, $F(G)$ for each gray level can be computed as below:

$$F(G) = \frac{Frequency\ of\ each\ gray\ level, G}{Total\ number\ of\ pixels\ in\ an\ image, N_p}$$

Based on the definition of $F(G)$ and knowing that number of gray level is represented by N_g, the formula for the features is described in Table 2.

Table 2. List of statistical measures and its corresponding formula.

Statistical measures	Formula
Mean	$Mean = \sum_{G=0}^{N_g-1} G \times F(G)$
Variance	$Var = \sum_{G=0}^{N_g-1} (G - mean)^2 \times F(G)$
Skewness	$S = \frac{\sum_{G=0}^{N_g-1} (G - mean)^3 \times F(G)}{Std^3}$
Kurtosis	$K = \frac{\sum_{G=0}^{N_g-1} (G - mean)^4 \times F(G)}{Std^4} - 3$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

Energy	$Energy = \frac{\sum_{G=0}^{N_g-1} F(G) \times F(G)}{{N_p}^2}$
Entropy	$Entropy = -\sum_{G=0}^{N_g-1} F(G) \times \log_2 F(G)$

2.3.2 GLCM

Developed by [18], GLCM is an excellent representation of image properties related to second order statistics [19].Nine textural features are used in this study. The following equations define these features. Let $P(i,j)$ be (i,j) entry for the GLCM while $p(i,j)$ be (i,j) entry for the normalized GLCM. Suppose that m and n denote the number of row and column of GLCM. The normalized GLCM entries, $p(i,j)$ can be expressed as:

$$p(i,j) = \frac{P(i,j)}{m \times n}$$

The mean and standard deviation for the rows and columns of GLCM are:

$$\mu_x = \sum_{i=1}^{m}\sum_{j=1}^{n} i.p(i,j), \mu_y = \sum_{i=1}^{m}\sum_{j=1}^{n} j.p(i,j)$$

$$\sigma_x = \sum_{i=1}^{m}\sum_{j=1}^{n} (i-\mu_x)^2.p(i,j), \sigma_y = \sum_{i=1}^{m}\sum_{j=1}^{n} (j-\mu_y)^2.p(i,j)$$

The features are described in Table 3.

Table 3. List of GLCM features and its corresponding formula.

GLCM features	Formula
Energy	$Energy = \sum_{i=1}^{m}\sum_{j=1}^{n} (p(i,j))^2$
Entropy	$Entropy = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(i,j).\log(p(i,j))$
Correlation	$Correlation = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(i.j).p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$
Homogeneity	$Homogeneity = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p(i,j)}{1+\lvert i-j\rvert}$
Absolute value	$Absolute\ value = \sum_{i=1}^{m}\sum_{j=1}^{n} \lvert i-j\rvert.p(i,j)$
Inertia	$Inertia = \sum_{i=1}^{m}\sum_{j=1}^{n} (i-j)^2.p(i,j)$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

Inverse difference	$Inverse\ difference = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{p(i,j)}{1+(i-j)^2}$
Maximum probability	$Maximum\ probability = \arg\max_{i,j} p(i,j)$
Autocorrelation	$Autocorrelation = \sum_{i=1}^{m}\sum_{j=1}^{n}(i.j).p(i,j)$

2.3.3 GLRLM

The frequencies of the gray level run length and its corresponding gray level that occur in digital image matrix is used to construct GLRLM. Elements in GLRLM is represented by $r(i,j)$, where i denotes the index for gray level whereas j denotes index for the run length. Before defining the textural features, let N_g be the number of gray levels in the image, N_r be the run length, m is the number of row and n is the number of column of original image. GLRLM features are described in Table 4.

Table 4. List of GLRLM features and its corresponding formula.

GLRLM features	Formula
Short run emphasis	$SRE = \frac{\sum_i^{N_g}\sum_j^{N_r}\frac{r(i,j)}{j^2}}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$
Long run emphasis	$LRE = \frac{\sum_i^{N_g}\sum_j^{N_r} j^2.r(i,j)}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$
Gray level distribution	$GLD = \frac{\sum_i^{N_g}(\sum_j^{N_r} r(i,j))^2}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$
Run length distribution	$RLD = \frac{\sum_j^{N_r}(\sum_i^{N_g} r(i,j))^2}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$
Run percentage	$RP = \frac{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}{m \times n}$
Low gray level run emphasis	$LGRE = \frac{\sum_i^{N_g}\sum_j^{N_r}\frac{r(i,j)}{i^2}}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$
High gray level run emphasis	$HGRE = \frac{\sum_i^{N_g}\sum_j^{N_r} i^2.r(i,j)}{\sum_i^{N_g}\sum_j^{N_r} r(i,j)}$

2.3.4 HOG

HOG proposed by [20] is a local feature descriptor that is able to capture edge gradient structure with high tolerance to local geometric and photometric (illumination) transformation [21] and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

simultaneously maintain high selectivity [22]. These features is well known for preserving local second-order interaction between pixels [23].

2.3.5 LBP

A grayscale and rotation invariant feature, known as local binary pattern (LBP) was introduced by [24]. LBP is a simple yet efficient operator in depicting local image pattern and has been utilized as features for various application [25].Aside from the implementation of original LBP, an improved version of locally rotation invariant LBP [26] is applied for better tradeoff between discriminative power and robustness.

2.3.6 DWPT

Unlike spectral analysis, e.g. Fourier transform (FT) that represents signal as sum of sinusoids, wavelet transform decomposes signal into basic wavelet functions of different scales in time-frequency domain. From the perspective of computational complexity, FT is $O(\frac{1}{2}n\log_2 n)$, while discrete wavelet transform (DWT) is just $O(n)$[27]. The inherent multiresolution property of wavelet transform is the main reason it is an excellent feature extraction method [28]. Essentially, the dyadic DWT can be implemented through filter tree algorithm. It is worth noticing that only approximation components will be extracted at each level of decomposition in DWT.

DWPT, being an extension to DWT allows all nodes in the tree structure to split [29]. In other word, not only approximation coefficients, details components are decomposed to form full binary tree as shown in Fig. 3. Three level of decomposition and Haar wavelet is chosen for this experiment. As shown in Fig. 3, The first series of analysis filter bank (low pass filter and high pass filter) is performed along the columns of the image while the second series is implemented along the rows of image or vice versa. LL, LH, HL and HH are subbands that capture approximation, horizontal edges, vertical edges and diagonal edges respectively.

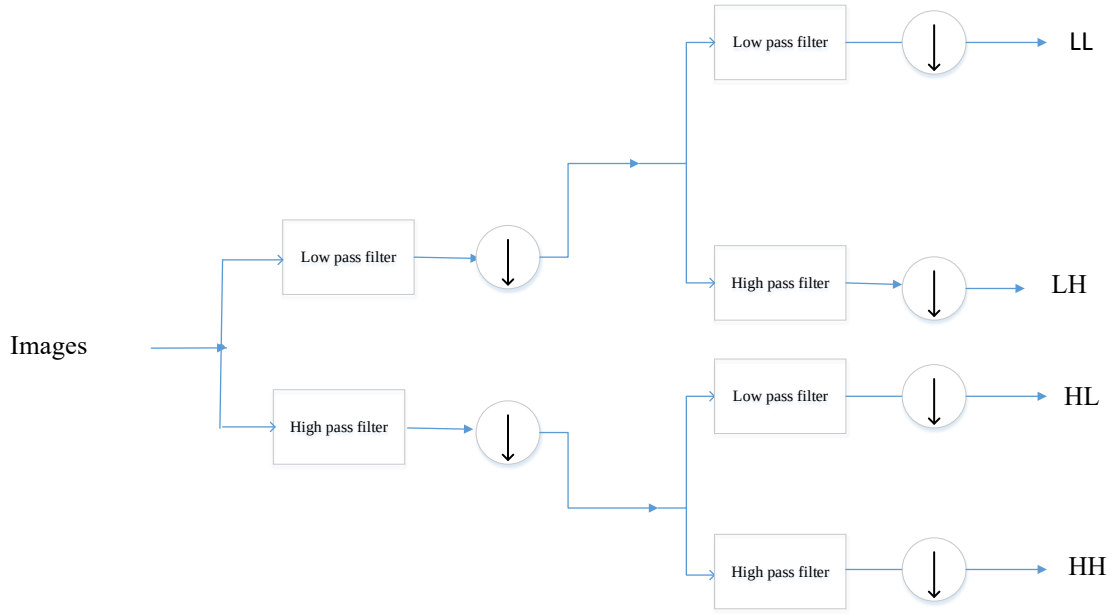


Fig. 3. Schematic diagram of single level of 2D DPWT decomposition.

2.4. Feature reduction/Feature subset selection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

Feature reduction, also known as dimensionality reduction is a process of transforming original attribute space into lower dimensional space, before they are input to classification models. Feature subset selection involves removal of irrelevant and redundant features. The objectives of these processes are to preserve informative and relevant features and get rid of irrelevant and redundant attributes, which can ultimately lead to improvement of classifiers' performance (generalization capability and convergence speed), simpler and more understandable models, and reduced storage requirement. In this paper, some widely used feature transform methods like Principal Component Analysis (**PCA**), Linear Discriminant Analysis (**LDA**), Partial Least Square (**PLS**) and Independent Component Analysis (**ICA**) [30] are employed. In addition, three filter feature selection method, namely **Relief**, Discriminant Least Square Regression (**DLSR)** [31] and Sequence Forward Search based on LW (**SFS-LW)** [32] were implemented. Essentially, these 2 techniques are different. Feature transform approach project the original data dimension into lower dimensional spaces through linear combination of original features, while feature selection seeks to choose subsets of features that is highly relevant to targets, or class labels [33].

2.5. Classification models

Classification models attempt to derive relationship between the set of input variables (normally in the form of input vector) and its corresponding categorical target/label. In this research, the classifiers are going to deal with binary classification problem (identify and distinguish neoplastic brain MRI from normal brain MRI). The classifiers employed in this study are logistic regression (LR), Naïve-Bayes (NB) classifier, kNN classifier, weighted kNN [34], artificial neural network (ANN), support vector machine (SVM), and extreme learning machine (ELM) [35]. Table 5 presents the implementation of the classification algorithms mentioned above.

Table 5. Implementation details of various classifiers.

Classification models	Implementation details
kNN	10-fold cross validation is used to determine optimum parameter, k (number of nearest neighbors) in the range from 1 to 100. Euclidean distance measures is used.
Weighted kNN	Inversion kernel function is to calculate the weights of nearest neighbors.
NB	Gaussian probability distribution function is used to quantify the class conditional probability of each features.
LR	Batch gradient descent algorithm is used to update the model coefficients.
ANN	Double layer hidden nodes: first one consists of 50, while second one consists of 30. Learning algorithm: batch gradient descent.
SVM	Two SVMs: linear and RBF. The sigma and C parameters in RBF SVM are optimized by grid search method [36].
ELM	Hidden nodes of 50 and 150 are used throughout the experiment. Optimum C parameters is found using 10-fold cross-validation method in the range of $\{2^{-24}, 2^{-23}, \dots, 2^{25}\}$[37].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

2.6. Performance evaluation and analysis

In order to evaluate the performance of each machine learning approaches applied, 6 widely accepted performance measures are utilized, including test accuracy (%), sensitivity (%), and specificity (%), AUC, training time (s) and test time (s). Since different schemes of feature reduction and classifiers will be experimented for 30 times using different sets of training data, in other words 30 different models will be trained, the mean and standard deviation of performance metrics can be computed for unbiased analysis and comparison among methods employed. Table 6 describes the performance measures and its definitions.

Different types of feature reduction and classification algorithms will then be applied. Finite size of feature subsets under each dimensionality reduction method will be experimented. Each experiment will be conducted 30 times using stratified sampling strategy for unbiased statistical analysis. With sufficiently large sample size, most distributions can be approximated as normal distribution according to Central Limit Theorem. Confidence intervals based on studentized method can be constructed after computation of standard error. However, this kind of comparison might be prone to type I error according to [38]. Fig. 4 shows the schematic diagram of the research framework.

2.7. Identification of relevant features

Several features are identified empirically by examining the change in accuracies according to number of features. The attributes are considered as relevant if the set of features added lead to significant improvement in accuracy. It should be noted that this empirical method is greedy and do not take between features interaction into account. Addition of features that leads to spike in accuracy is identified the *goodness of split* of each identified attribute is computed. Goodness of split is a measure used by classification and regression tree (CART) to choose which attribute to split. The first three features with highest goodness of split are selected for construction of decision tree. Further details were discussed in section 3.3.

2.8. Construction of decision tree and classification rule

Blind use of complex and black-box predictive models in high stake circumstances had led to serious consequences [39]. This scenario may be induced by false widespread belief of accuracy and interpretability trade-off among researchers which suggests that more complex model is more accurate [40].

These recent study and findings have motivated and led to this extra step of determining relevant feature subset and constructing decision tree and rule-based classifiers. This not only achieve the objective of building interpretable models but also contributes to knowledge discovery as the mechanisms of predictive models in predicting the outcomes (normal or tumorous) is better understood.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

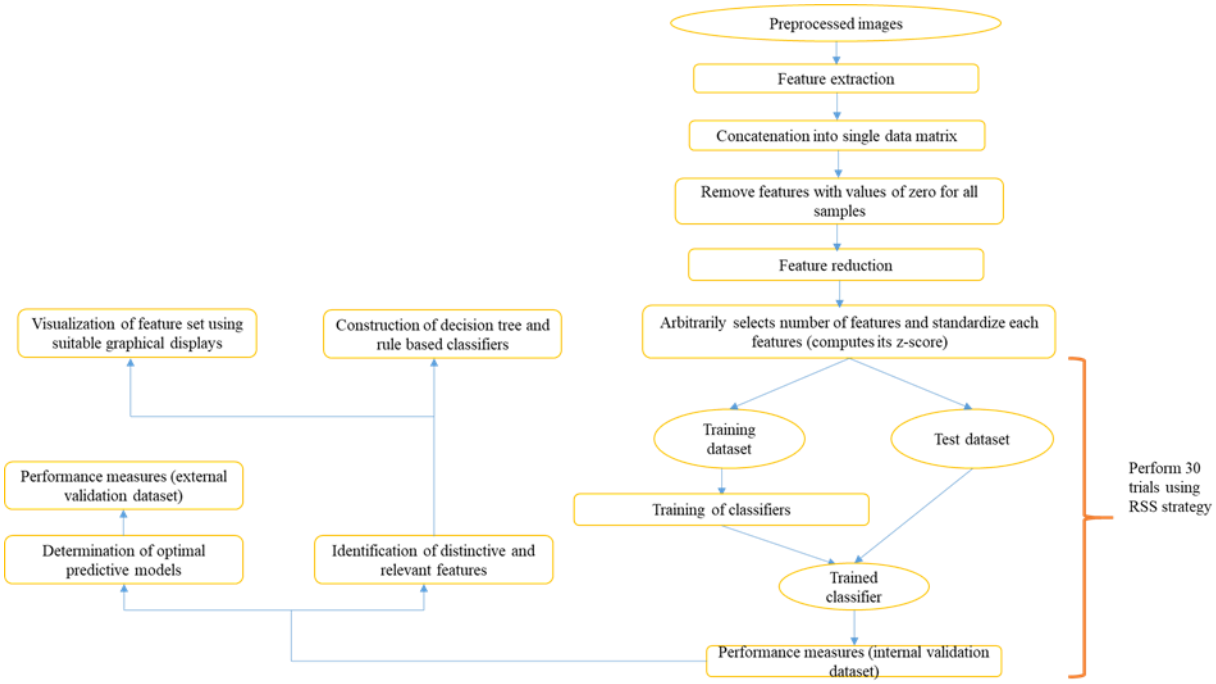


Fig. 4. Process flow of experiments conducted.

Table 6. Performance measures and its definitions.

Performance measures	Definition
Accuracy	Overall efficiency and generalizability of classifier [41].
Sensitivity	Ability to correctly identify positive samples (necrotic brain MR images).
Specificity	Ability to correctly identify negative samples (normal brain MR images).
AUC	Equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than randomly chosen negative instance [8].
Training time	Quantify the convergence speed of training.
Test time	Quantify the speed of generation of test data output labels

3. Results and Discussions

3.1 *Comparison among classifiers under different feature reduction schemes and number of features*

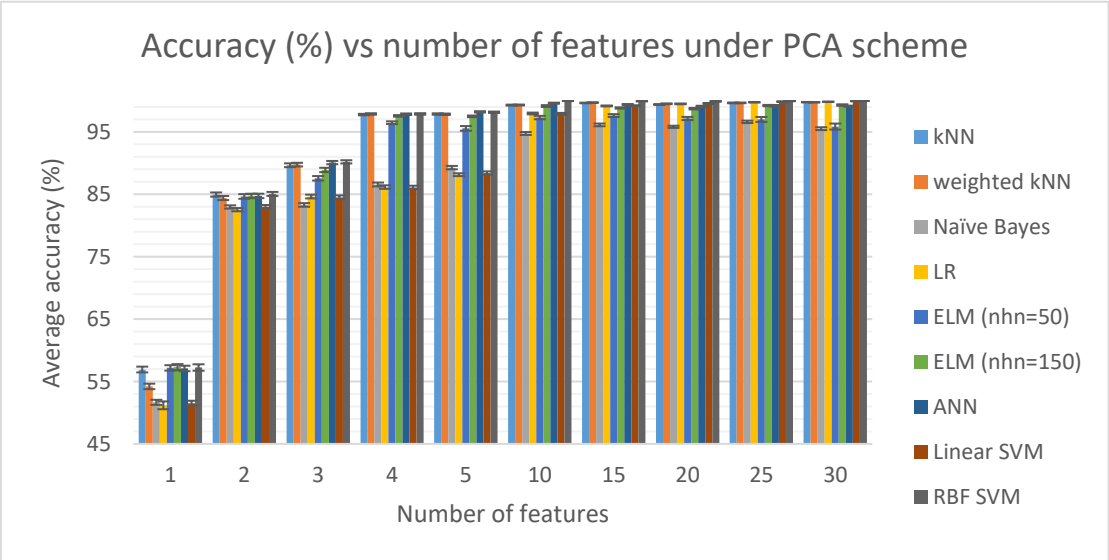
As mentioned in section 2.1, dataset 1, 2, 4 and 5 are employed in experiment of this section. For each 30 trials, the dataset is partitioned into 70% training data and remaining 30% as test data using random stratified sampling (RSS) strategy. Seven dimensionality reduction methods are employed in this research and their performances will be evaluated in terms of the number of features and types of classifiers. Here, we will examine the performance of classifiers from the perspective of each dimensionality reduction techniques:

- Under PCA scheme, radial basis function (RBF) SVM achieves the highest average accuracy (99.969%±0.015) when the number of attributes is 30. Nevertheless, PCA+RBF SVM with 5 features also achieved very high accuracy of (99.96%±0.017). With the increase in the number

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

of predictors, the accuracy of classifiers increase and become stable beyond 20 features.

- Under PLS scheme, the accuracy of classifiers improved with more feature input until the number of feature reaches 10. The highest accuracy $(99.998\% \pm 0.005)$ is achieved by RBF SVM with 10 features. It is worth noting that all classifiers except NB classifier achieve accuracy of more than 99% when number of feature increase beyond 10.
- Under LDA scheme, the performance of classification models varied greatly across different number of features. The accuracy of both ELM with number of hidden nodes 50 and 150 increase dramatically up to $84.234\% \pm 0.817$ and $93.84\% \pm 0.225$ with just 2 attributes. Spike in accuracy of kNN, weighted kNN and RBF SVM when number of features increase to 20. With 45 attributes, linear SVM, kNN and weighted kNN experience improvement of accuracy. ***Perfect classification (accuracy of 100%) are achieved by ELM with 150 hidden neurons when number of features are 15, 30, 40, 45 and 55.*** On the other hand, the accuracy of NB classifier decrease when number of features increase beyond 25 which imply overfitting. It is worth noting here that NB classifiers display strong bias towards classifying the test instance to be negative (normal) as suggested by the decrease in sensitivity and increase in specificity.
- Under ICA scheme, RBF SVM achieves the highest average accuracy ($99.891\% \pm 0.029$) when the number of attributes is 30. All classifiers have their accuracy improve with more features included until 25 features where the accuracy become stable.
- Under DLSR feature selection scheme, all the classifiers attain classification accuracy of more than 99% when the number of feature reaches 5. ***Perfect classification are obtained by LR with number of features are 45 and 50, linear SVM with number of features are 35, 40, 45 and 50 and RBF SVM with number of features are 35, 45 and 50.***
- Under Relief feature selection scheme, the highest accuracy $(99.993\% \pm 0.008)$ is achieved by RBF SVM with 30 predictors. All classifiers, except NB classifier obtain more than 99% in accuracy when number of feature reaches 15.
- Under SFS-LW feature selection scheme, accuracy all classifiers except Naïve Bayes classifier increase dramatically when number of features reach 11, 12 and 13. The optimal classifier is RBF SVM with accuracy of $99.962\% \pm 0.019$ with 40 features.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

Fig. 5. Classification accuracy of PCA + different classifiers with different number of features. The error bars are plotted using values of: average accuracy $\pm \frac{t(\text{standard deviation})}{\sqrt{n}}$, where t represents t multiplier for student's t distribution. The value of t is 2.0452 (degree of freedom=n-1, probability=0.05). n represents the number of samples. The ranges of accuracy represent 95% of confidence intervals. It should be noted that this CI can be misleading due to variation in the sampling of training data and should not be used for significance test and comparison.

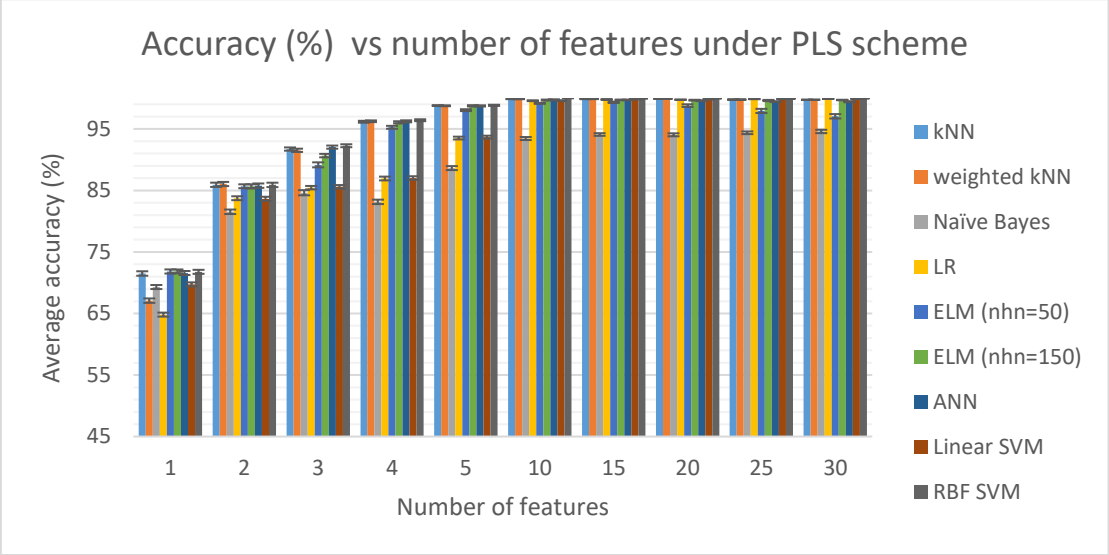


Fig. 6. Classification accuracy of PLS + different classifiers with different number of features.

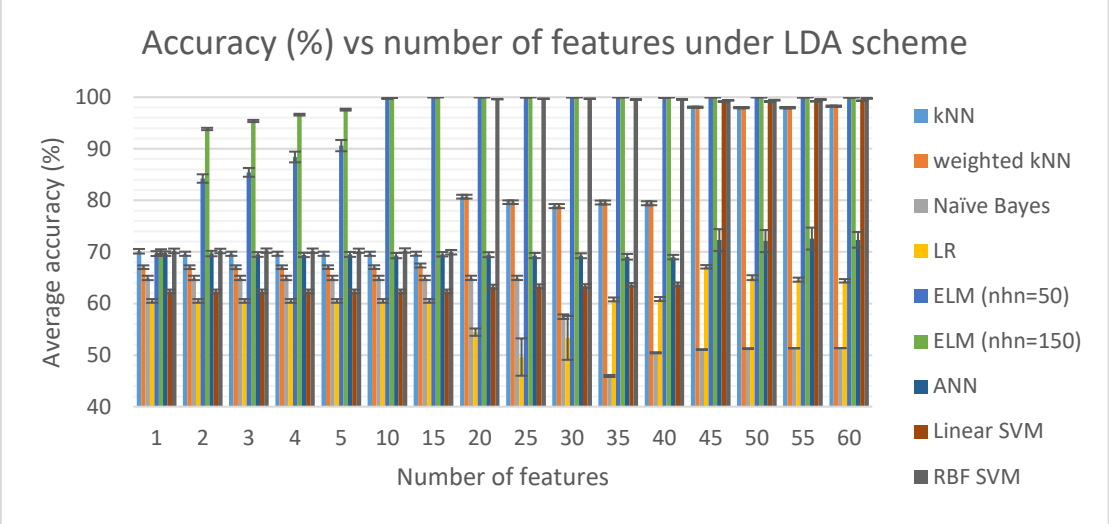


Fig. 7. Classification accuracy of LDA + different classifiers with different number of features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

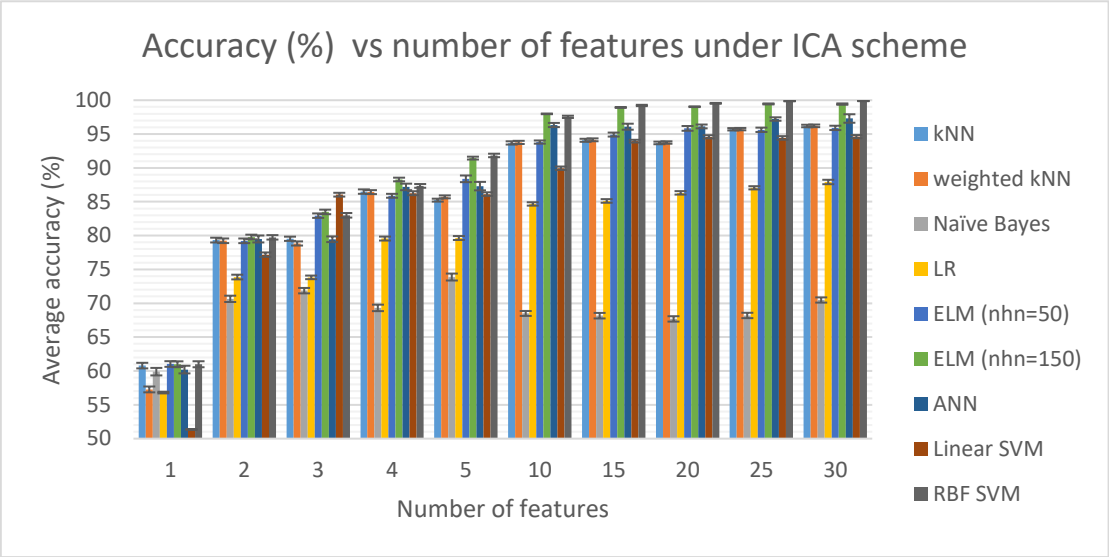


Fig. 8. Classification accuracy of ICA + different classifiers with different number of features.

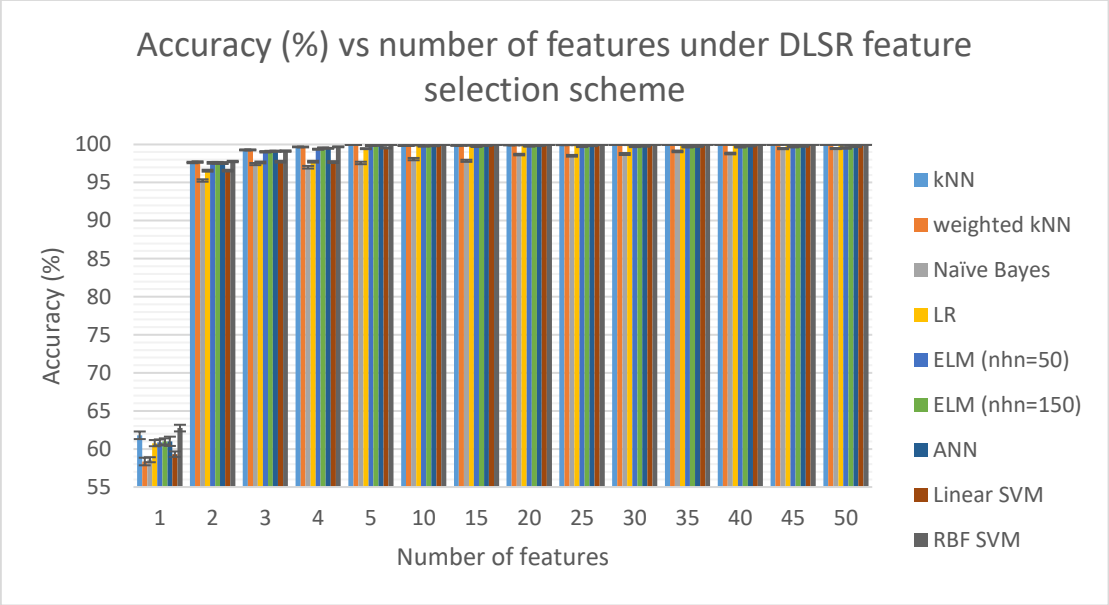


Fig. 9. Classification accuracy of DLSR + different classifiers with different number of features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

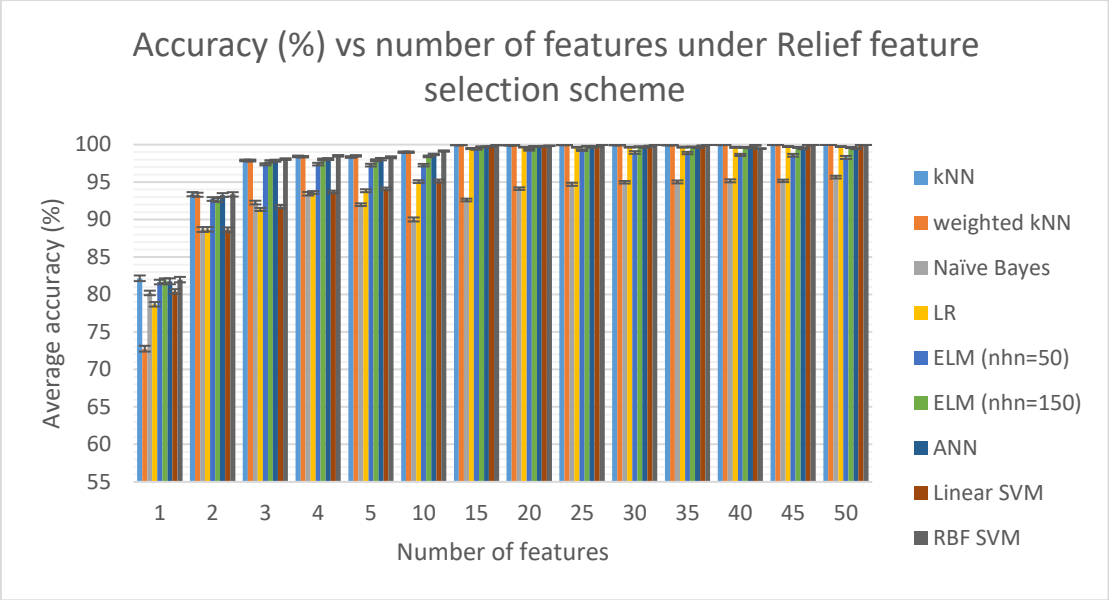


Fig. 10. Classification accuracy of Relief+ different classifiers with different number of features.

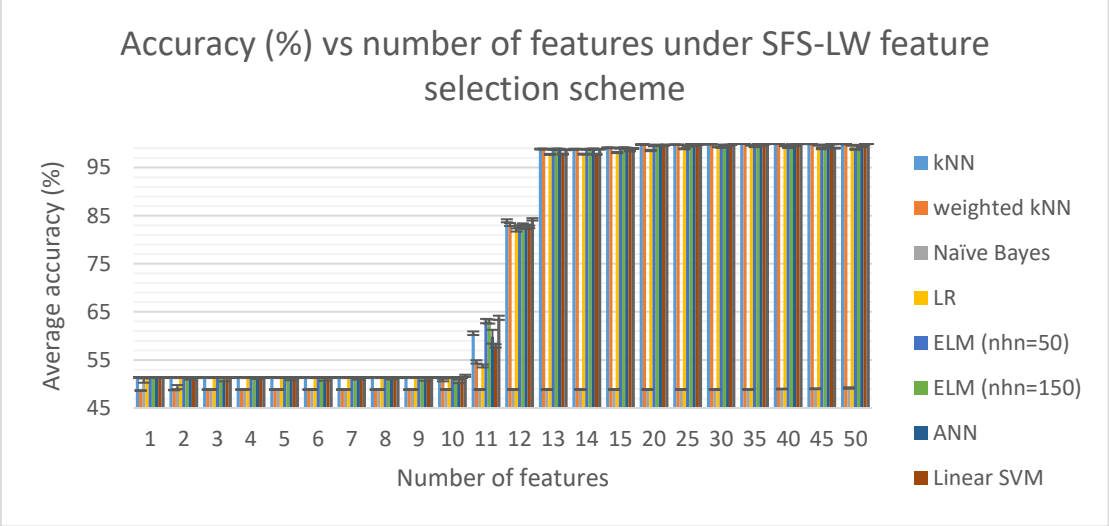


Fig. 11. Classification accuracy of SFS-LW+ different classifiers with different number of features.

3.2 *Performance evaluation using external validation set*

As mentioned in section 2.1, dataset 1, 2, 4 and 5 will serve as training dataset and dataset 3 and 6 will serve as test dataset. The purpose of this extra experimentation is to gain insight of the generalizability of the machine learning pipelines that had achieved perfect classification as pointed out in section 3.1. In addition to that, we manipulate external validation set (dataset 3 and 6) to be balanced (100 tumorous cases and 100 normal cases) and imbalanced (100 tumorous cases and 900 normal cases) so that each perfect classifiers can be evaluated under both balanced and skewed class distribution. According to Table 7 and 8, LDA+ELM with 55 attributes achieved the highest accuracy: 0.975 (balanced class distribution) and 0.995 (imbalanced class distribution). Specificity of 1 suggest that the classifier is expert in identifying normal images. Sensitivity of 0.95 indicates that there are 5 tumorous test instances misclassified as normal. On the other hand, all other machine learning schemes show discrepancy in performances for both balanced and imbalanced test set, in which the performance deteriorate when test

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

with imbalanced dataset. In other words, the learning algorithms misclassify the normal images as tumorous with newly added normal cases in the dataset.

Table 7. Performance measures of optimal machine learning pipeline under balanced test dataset (100 tumorous cases and 100 normal cases).

Dimensionality reduction methods	Classifiers	number of features	Performance measures						
			Accuracy	AUC	Sensitivity	Specificity	Precision	Training time	test time
LDA	ELM (number of hidden nodes=150)	15	0.935	0.990	1.000	0.870	0.885	0.027	0.001
		30	0.945	0.962	0.980	0.910	0.916	1.251	0.097
		40	0.940	0.970	0.980	0.900	0.907	0.029	0.001
		45	0.920	0.929	0.930	0.910	0.912	0.026	0.001
		55	**0.975**	**0.989**	**0.950**	**1.000**	**1.000**	**0.030**	**0.001**
DLSR	LR	45	0.935	0.997	1.000	0.870	0.885	0.209	0.012
		50	0.910	0.979	0.990	0.830	0.853	0.129	0.000
	linear SVM	35	0.950	1.000	1.000	0.900	0.909	0.336	0.002
		40	0.965	1.000	1.000	0.930	0.935	0.247	0.003
		45	0.965	1.000	1.000	0.930	0.935	3.094	0.252
		50	0.965	1.000	1.000	0.930	0.935	0.180	0.001
	RBF SVM	35	0.895	0.936	1.000	0.790	0.826	0.190	0.008
		45	0.915	0.970	1.000	0.830	0.855	0.224	0.005
		50	0.850	0.958	0.880	0.820	0.830	0.197	0.003

Table 8. Performance measures of optimal machine learning pipeline under imbalanced test dataset (100 tumorous cases and 900 normal cases).

Dimensionality reduction methods	Classifiers	number of features	Performance measures						
			Accuracy	AUC	Sensitivity	Specificity	Precision	Training time	test time
LDA	ELM (number of hidden nodes=150)	15	0.643	0.966	1.000	0.603	0.219	0.047	0.006
		30	0.649	0.897	0.980	0.612	0.219	0.028	0.005
		40	0.639	0.884	0.980	0.601	0.214	0.028	0.005
		45	0.641	0.856	0.930	0.609	0.209	0.024	0.005
		55	**0.995**	**0.988**	**0.950**	**1.000**	**1.000**	**0.030**	**0.006**
DLSR	LR	45	0.853	0.990	1.000	0.837	0.405	0.137	0.000
		50	0.804	0.974	0.990	0.783	0.337	0.129	0.000
	linear SVM	35	0.880	1.000	1.000	0.867	0.455	0.338	0.002
		40	0.926	1.000	1.000	0.918	0.575	0.190	0.002
		45	0.895	1.000	1.000	0.883	0.488	6.466	0.363
		50	0.895	1.000	1.000	0.883	0.488	0.180	0.002
	RBF SVM	35	0.697	0.882	1.000	0.663	0.248	0.786	0.959
		45	0.745	0.918	1.000	0.717	0.282	0.209	0.007
		50	0.727	0.905	0.880	0.710	0.252	0.212	0.013

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

3.3 *Identification of relevant features subset*

Several features are identified empirically by examining the change in accuracy in Fig. 9-11 under DLSR, Relief, and SFS-LW feature selection methods. Table 9 shows the first 15 features of the 3 feature selection schemes. Several significant features were determined based on the improvement in accuracy scores after addition of particular feature as shown in Fig. 12. There are 10 features being identified that can potentially provide information in classification of normal and tumorous image samples. Then, MATLAB function predictorImportance was utilized to acquire the predictors' importance estimate. The results are shown in Fig. 13. Three of the most important features have their indices attached with asterisk in Table 9. The details of these 3 features are listed in Table 10. In addition, we also visualize the most informative features in Fig. 14-16. As can be seen in Fig. 14, 2 peaks observed in the kernel density plot strongly suggest bimodal distribution. In Fig. 15 and Fig. 16, most of the normal and tumorous samples represented by the features occupy different regions in the dimensional space. This graphic displays have demonstrated the discriminating power of the selected features.

Table 9. First 15 features selected by the proposed feature selection techniques. The bold numbers are indices of discriminative features.

Indices of features	DLSR	Relief	SFS-LW
1	**1922**	**1931**	1865
2	**61***	**1939***	1859
3	**1909***	**302**	1915
4	**1928**	303	1903
5	**60**	309	1896
6	1884	310	1870
7	55	284	1840
8	427	545	1897
9	1846	1928	1921
10	1522	546	1859
11	59	1930	**1835**
12	1876	1922	**1881**
13	991	1909	**61**
14	1931	285	1909
15	135	1933	1939

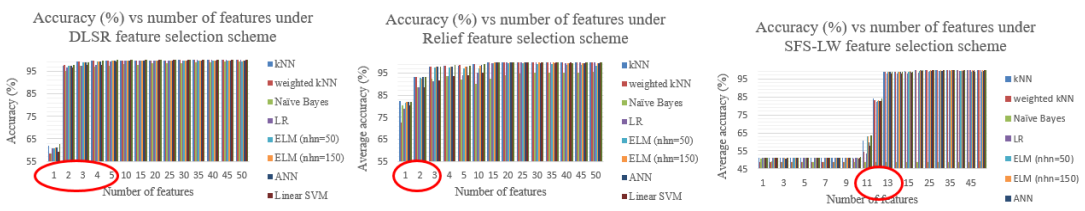


Fig. 12. Manual determination of significant features (shown in bold in Table 9).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

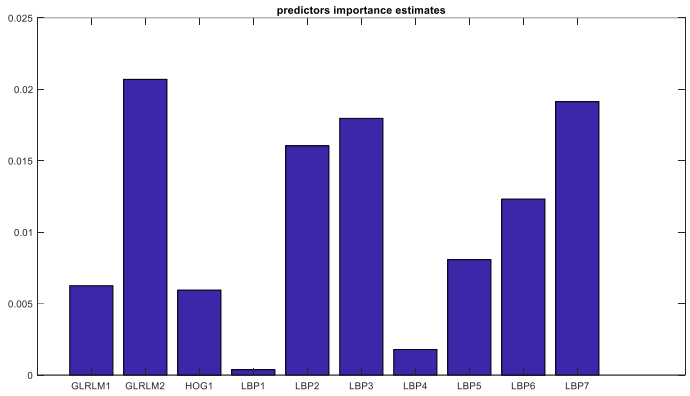


Fig. 13. Bar chart depicting the estimates of predictors' importances.

Table 10. The most important features arranged in descending order: 61th feature, 1939th feature, 1909th feature.

	Description
61th feature	Gray level run length non-uniformity at angle of 90°.
1939th feature	Probability of rotation invariant local binary pattern that corresponds to uniformity measures more than 2 in an image sample.
1909th feature	Probability of local binary pattern with $LBP_{8,1}$ $value\ of$ 152.

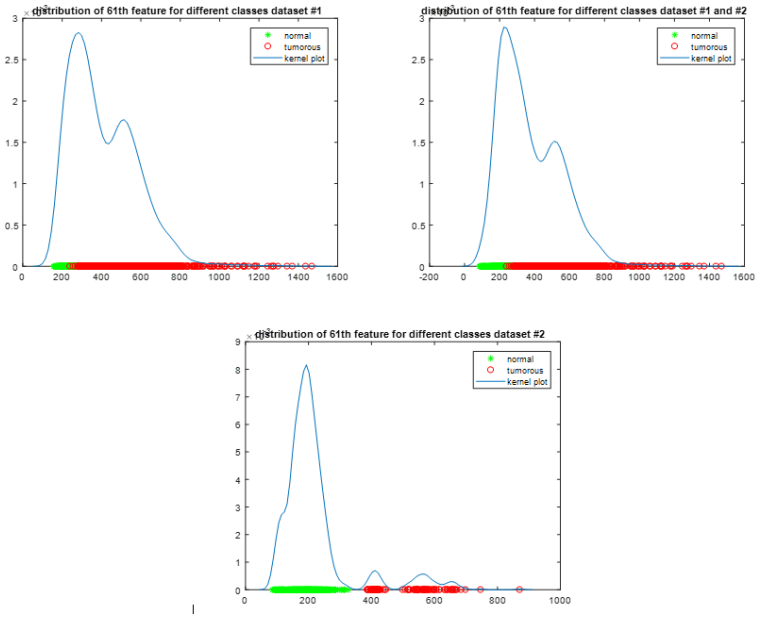


Fig. 14. Kernel plots of 61th feature.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

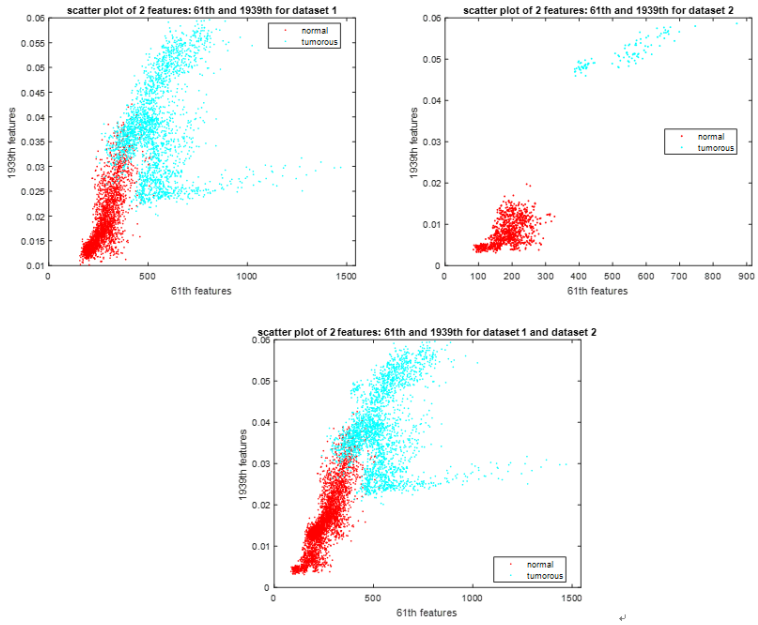


Fig. 15. 2D scatter plots for 2 features (61th and 1939th).

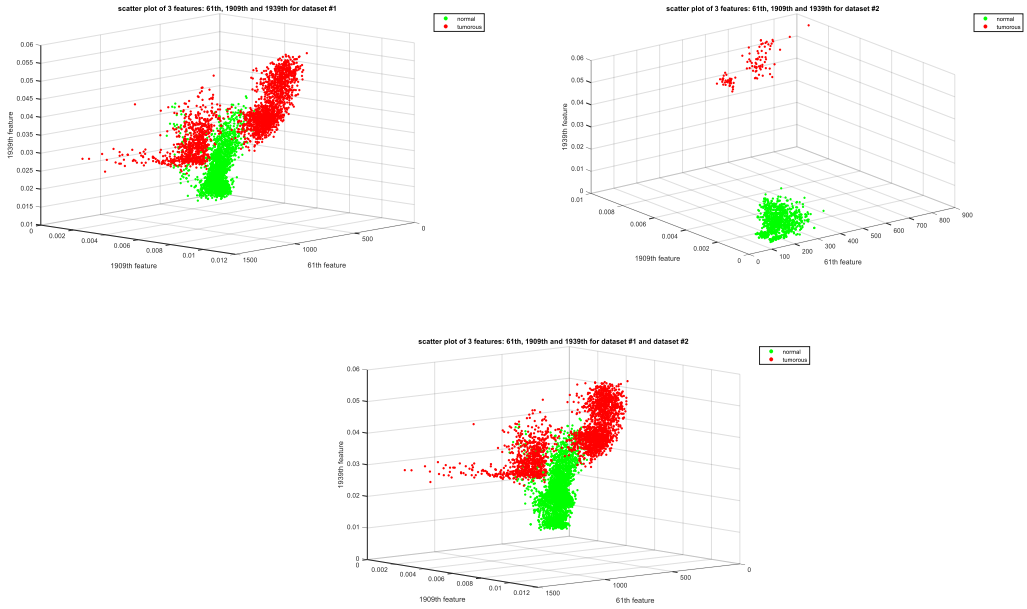


Fig. 16. 3D scatter plots for 3 features (61th, 1909th and 1903th).

3.4 Construction of decision tree and set of rules

In this section, **all downloaded dataset** was used to train 10-fold cross validated decision tree classifier. The changes in performance measures when the relevant features are gradually added to input data are shown in Table 11. Accuracy as much as 0.988 is attained when trained with input data of 3 attributes. The high sensitivity (low false negative rate) is crucial as cost of false negatives is higher than the cost of false positives prediction[42]. In order to track down the best models comprising of the 3 features, we look into the 10 cross-validated decision tree models. The optimal decision tree models trained with data of 3 features are shown in Fig. 17 and its performance shown in Table 12.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

Table 11. Average performance measures of cross-validated decision tree models with different number of features.

Performance measures	1 features (61th)		2 features (61th and 1939th)		3 features (61th, 1909th, and 1939th)	
	Average	standard deviation	Average	standard deviation	Average	standard deviation
Accuracy	0.941	0.008	0.966	0.009	**0.988**	0.005
AUC	0.938	0.008	0.984	0.004	0.991	0.005
Sensitivity	0.922	0.013	0.980	0.004	0.996	0.005
Specificity	0.955	0.009	0.956	0.013	0.982	0.009
Precision	0.936	0.011	0.941	0.016	0.976	0.011
false positive rate	0.045	0.009	0.044	0.013	0.018	0.009
miss rate	0.078	0.013	0.020	0.004	0.004	0.005

Table 12. Performance measures of each cross-validated decision tree models when data of 3 features are applied.

Models	Accuracy	AUC	Sensitivity	Specificity	Precision	False positive rate	False negative rate
1	0.9912	0.9926	1.0000	0.9848	0.9793	0.0152	0.0000
2	0.9806	0.9797	1.0000	0.9667	0.9558	0.0333	0.0000
3	0.9842	0.9877	1.0000	0.9727	0.9636	0.0273	0.0000
4	**0.9947**	**0.9955**	**1.0000**	**0.9909**	**0.9876**	**0.0091**	**0.0000**
5	0.9912	0.9939	1.0000	0.9848	0.9794	0.0152	0.0000
6	0.9859	0.9897	1.0000	0.9758	0.9675	0.0242	0.0000
7	0.9929	0.9975	0.9916	0.9939	0.9916	0.0061	0.0084
8	0.9841	0.9869	0.9916	0.9788	0.9711	0.0212	0.0084
9	0.9859	0.9898	0.9916	0.9818	0.9751	0.0182	0.0084
10	0.9894	0.9958	0.9873	0.9909	0.9873	0.0091	0.0127

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

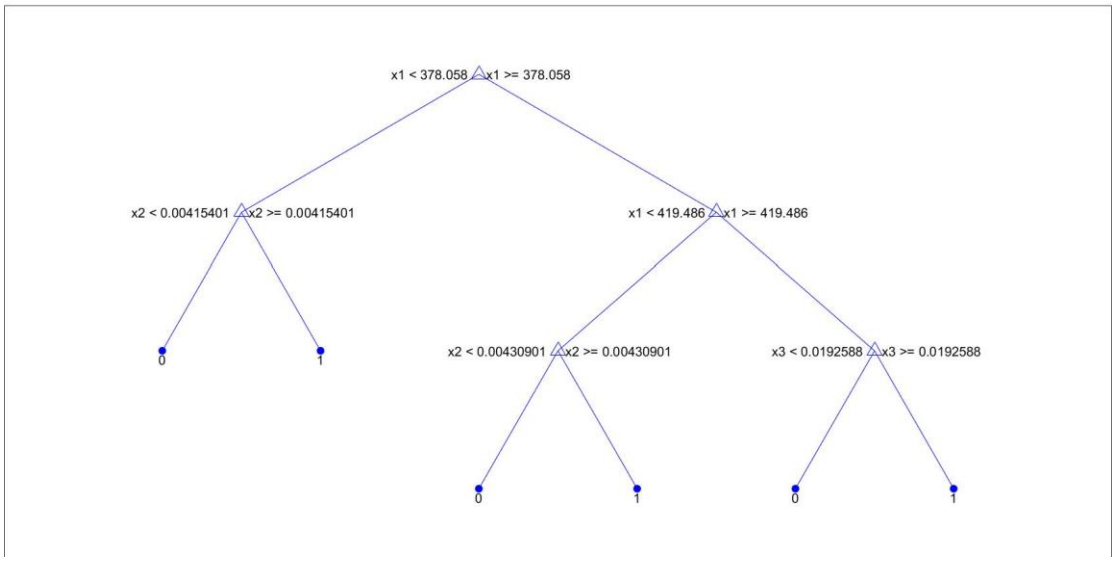


Fig. 17. Decision tree model 4. x1, x2 and x3 denote features 61th, 1909th and 1939th respectively.

The decision tree above can be converted to rules by following each path from parent nodes to every single leaf nodes as shown below:

$$If\ \boldsymbol{x_{61}} < 378.058\ and\ x_{1909} \geq 0.00415401, then\ the\ sample\ x\ is\ tumorous$$

$$If\ x_{61} \geq 419.486\ and\ x_{1939} \geq 0.0192588, then\ the\ sample\ x\ is\ tumorous$$

$$If\ 378.058 \leq x_{61} < 419.486\ and\ x_{1909} \geq 0.00430901, then\ the\ sample\ is\ tumorous$$

$$otherwise, normal.$$

The classification rules can be visualized as in Fig 18.

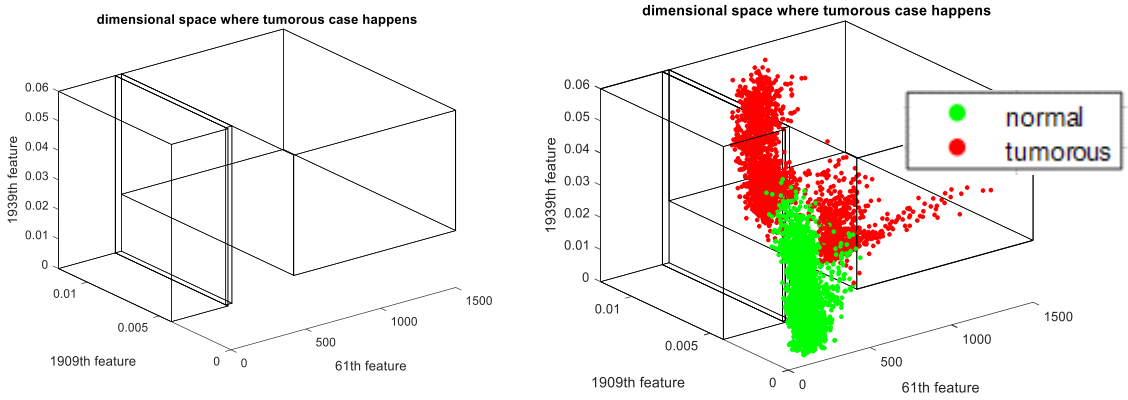


Fig. 18. (Left) The space enclosed by the cuboids are where tumorous instances lie. (Right) Both normal and tumorous points are plotted on top of the 3D plot for better visualization purpose.

4. Discussion

We will arrange discussion of our results according to order in Section 3 paragraph by paragraph.

To summarize, there are some important findings discovered from Section 3.1:

a. For most learning algorithms, the accuracy of classifiers tends to improve and become stable after a certain number of attributes as the number of features increase. The performance of classifiers can be affected by the ability of classification models to deal with large number of attributes. Classifier like RBF SVM that can project low dimensional space to high dimensional

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

space can achieve better classification accuracy with fewer features, and thus resilient to overfitting. On the other hand, Naïve Bayes performs poorly across all feature reduction methods. This may be attributed to *violation of independence among attributes* and Gaussian probability distribution might not be *good estimation of probability distribution* of the features.

b. kNN and weighted kNN shows similar trends in their accuracy, however the test time for weighted kNN is significantly higher compared to kNN. In weighted kNN, the weight of each neighbor has to be evaluated to determine class of new test instance.
c. The average accuracy for ANN is generally lower compared to both ELM and SVM, particularly when number of predictors increases, which often come with higher requirement of training time. This may be attributed to overfitting.
d. RBF SVM is very robust and achieves relatively high test accuracy in all feature reduction schemes compared to other classifiers, except under LDA schemes where ELM outperforms SVM. Nonetheless, the parameters C and σ is nontrivial and has to estimated using cross validated grid search method, which is computationally intensive.
e. All the dimensionality reduction techniques proposed is feasible with high accuracy achieved with selection of appropriate number of features and classifiers.
f. Classifiers under PCA and PLS scheme requires less features to achieve high accuracy if compared with LDA and ICA scheme.
g. Number of features has to be carefully selected under LDA and ICA scheme as the accuracy can vary substantially with the change in number of attributes included as shown in Fig. 7 and Fig. 8.
h. DLSR feature selection is more superior to Relief and SFS-LW as classifiers fed with DLSR feature attain higher accuracy with fewer features. This means DLSR can select relevant and more compact feature subsets.
i. Training time of classifiers can be sorted in ascending order as follows: ELM < linear SVM < LR < RBF SVM < ANN. We do not take into account Gaussian NB classifier as the training phase only involves computation of mean and standard deviation of each features of training data.
j. The training time of RBF SVM with ICA features as input variables is also significantly higher. According to [43], the training time complexity of SVM is proportional to number of support vectors. Thus, it is believed that number of support vectors is higher in ICA feature space.
k. Test time of all classifiers stagnate regardless of number of features except for kNN, weighted kNN and NB. The calculation of Euclidean distance of nearest neighbors and posterior probability becomes more complex with inclusion of more features.
l. Since all of the dimensionality reduction methods are viable, we are more interested in feature selection scheme as the attributes under feature transform methods (e.g. PCA, PLS, LDA and ICA) loses its interpretability and the contribution of each original features are unknown [44]. By examining Fig. 9-11 carefully, we spotted a huge spike in accuracy when certain features are included. We suspect there is some relevant features in each of those spikes in accuracy and a few features that contribute to improvement in accuracy are selected.

In section 3.2, we further evaluate the optimal supervised classification models using dataset from different sources. Additionally, dataset was manipulated in term of proportion of tumorous and normal cases to assess the classifiers using balanced and imbalanced test data. Most machine learning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2020 doi:10.20944/preprints202007.0688.v1

paradigms evaluated perform differently under balanced and imbalanced test dataset, except LDA+ELM classifier with 55 features. This shows that the *generalization performance of classifiers will generally deteriorate when tested using data from other sources, in our case different MRI machine and parameter settings.* Thus, we suggest the *training data samples should be vast and diverse with relevant predictors included, which could only be achieved with extensive experimentation.*

From the results shown in Fig. 9-11, it is clear that certain attributes contribute to the improvement of classifiers. We further identified these features and evaluate their importances based on decrease in impurity measures (Gini index) due to split. Three of the most important attributes are selected. The reasons for the relevance of these features are listed below:

- Gray level run length non-uniformity measures the heterogeneity of gray level. Based on Fig. 13, high value of run length non-uniformity indicate tumorous cases and vice versa. Theoretically, high value of run length non-uniformity implies the runs are not equally distributed across length [45]. The authors suspects that the unequal distribution might arise from the presence of tumor where different run length of similar pixel intensity for some gray levels might occur within the tumor region.
- In line with what is reported in [24], good discrimination can be achieved with the occurrence probability of 'uniform' rotation invariant local binary patterns. In this research, the frequency (probability) of LBP pattern with uniformity measure of more than 2. This is probably caused by the heterogeneity at the boundary of tumor tissue which can result in higher values for this attribute as shown in Fig. 14.

In section 3.4, with the use of these relevant features, 10-fold cross-validated decision tree models was trained and achieved 98.8% accuracy. The determination of informative feature subsets is more critical than the utilization of complex, black-box model. The high discriminative power and relevance of this feature subset *suggest the feasibility to build simple classification rules with fairly high accuracy. Classification rules is favorable as it can serve as independent nuggets of knowledge [46]. This model is not only intuitive and can be interpreted easily by end users*, *but can leads to knowledge discovery apart from making prediction on unseen data.*

5. Conclusions

The high resolution MR images has become one of the most popular neuroimaging tools nowadays, not just for research in brain anatomical structure, but can be useful in medical diagnosis (tumor detection). Machine learning is definitely a promising pathway in the development of automated CAD system. The first part of the experiment show high accuracy can be obtained for almost all machine learning pipelines with careful selection of number of predictors when *trained and tested with dataset of same sources*. We identified the machine learning schemes with perfect classification. Then, we evaluate these trained classifiers with independent test dataset of different source to further validate the performance of each approaches. Experimental results show that the optimal techniques is LDA+ELM (number of hidden neurons=150) with 55 features achieving (accuracy=97.5%, AUC=0.989, sensitivity=95% and specificity=100%) under balanced test dataset; (accuracy=99.5%, AUC=0.988, sensitivity=95% and specificity=100%). This particular model excels in predicting potential normal MR images. In addition, we also identified a few highly relevant features (1 from GLRLM and 2 from LBP) and visualized them graphically. With the data showing high separability in scatter plot, relatively easy-to-read 10 fold cross validated decision tree models are constructed by using both dataset #1 and #2 with the three features. The performances are as follow: accuracy=98.8%, AUC=99.1%, sensitivity=99.6% and

specificity=98.2%. Then, simple classification rule read directly off the decision tree are constructed. This highly intuitive classification rule sets which involves only 3 predictors can be easily embedded in any real time expert systems or automated CAD tools.

There are numerous possible directions for future work in this field. More advanced feature descriptors can be explored to search for the robust, repeatable and reproducible radiomic features in tumor diagnosis [47]. Less popular nonlinear dimensionality reduction methods like kPCA and locally linear embedding can also be investigated.

Abbreviations

MR: Magnetic resonance; CAD: Computer aided diagnosis; GLCM: Gray level co-occurrence matrix; GLRLM: Gray level run length matrix; HOG: Histograms of oriented gradients; LBP: Linear Binary Pattern; PCA: Principal component analysis; LDA: Linear Discriminant Analysis; PLS: Partial Least Square; ICA: Independent Component Analysis; DLSR: Discriminant Least Square Regression.

Declarations

Availability of data

The sources of data have been stated clearly in section 2.1.

Competing interests

N/A

Fundings

A special thanks to Universiti Teknologi Malaysia (UTM) for the opportunity to carry out the research and Ministry of Education (MOE) for financial support. This project was supported by Research University Grant [Vot Number: 19H03] initiated by UTM and MOE.

References

[1] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "Design and implementation of a computer-aided diagnosis system for brain tumor classification," in *2016 28th International Conference on Microelectronics (ICM)*, 2016, pp. 73-76.

[2] N. Nabizadeh and M. Kubat, "Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features," *Computers & Electrical Engineering,* vol. 45, pp. 286-301, 2015/07/01/ 2015.

[3] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martinez-Murcia, and I. for the Alzheimer's Disease Neuroimaging, "Automatic ROI Selection in Structural Brain MRI Using SOM 3D Projection," *PLOS ONE,* vol. 9, no. 4, p. e93851, 2014.

[4] H. Greenspan, B. v. Ginneken, and R. M. Summers, "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1153-1159, 2016.

[5] S. Iqbal, M. U. G. Khan, T. Saba, and A. Rehman, "Computer-assisted brain tumor type discrimination using magnetic resonance imaging features," *Biomedical Engineering Letters,* vol. 8, no. 1, pp. 5-28, 2018/02/01 2018.

[6] K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research - commentary," *Biomedical engineering online,* vol. 13, pp. 94-94, 2014.

[7] V. Kumar *et al.*, "Radiomics: the process and the challenges," (in eng), *Magnetic resonance imaging,* vol. 30, no. 9, pp. 1234-1248, 2012.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, no. 8, pp. 861-874, 2006/06/01/ 2006.

[9] J. Cheng *et al.*, "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," (in eng), *PLoS One,* vol. 10, no. 10, p. e0140381, 2015.

[10] J. Cheng *et al.*, "Retrieval of Brain Tumors by Adaptive Spatial Pooling and Fisher Vector Representation," (in eng), *PLoS One,* vol. 11, no. 6, p. e0157112, 2016.

[11] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Medical Image Analysis,* vol. 4, no. 1, pp. 43-55, 2000/03/01/ 2000.

[12] M. R. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, and R. Kikinis, "Automated Segmentation of MR Images of Brain Tumors," *Radiology,* vol. 218, no. 2, pp. 586-591, 2001.

[13] L. Mercier, R. F. Del Maestro, K. Petrecca, D. Araujo, C. Haegelen, and D. L. Collins, "Online database of clinical MR and ultrasound images of brain tumors," (in eng), *Med Phys,* vol. 39, no. 6, pp. 3253-61, Jun 2012.

[14] S. F. Eskildsen *et al.*, "BEaST: brain extraction based on nonlocal segmentation technique," (in eng), *Neuroimage,* vol. 59, no. 3, pp. 2362-73, Feb 1 2012.

[15] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable," (in eng), *IEEE transactions on medical imaging,* vol. 31, no. 2, pp. 153-163, 2012.

[16] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement," *Annals of Internal Medicine,* vol. 162, no. 1, pp. 55-63, 2015.

[17] J. Juntu, J. Sijbers, D. Van Dyck, and J. Gielen, "Bias Field Correction for MRI Images," in *Computer Recognition Systems*, Berlin, Heidelberg, 2005, pp. 543-551: Springer Berlin Heidelberg.

[18] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-3, no. 6, pp. 610-621, 1973.

[19] M. Partio, B. Cramariuc, M. Gabbouj, and A. Visa, *Rock texture retrieval using gray level co-occurrence matrix*. 2002.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893 vol. 1.

[21] M. Annalakshmi, S. M. M. Roomi, and A. S. Naveedh, "A hybrid technique for gender classification with SLBP and HOG features," *Cluster Computing,* 2018/01/30 2018.

[22] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," (in eng), *Neuron,* vol. 73, no. 3, pp. 415-34, Feb 9 2012.

[23] H. Bristow and S. Lucey, *Why do linear SVMs trained on HOG features perform so well?* 2014.

[24] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 971-987, 2002.

[25] Z. Guo, L. Zhang, and D. Zhang, "A Completed Modeling of Local Binary Pattern Operator for Texture Classification," *IEEE Transactions on Image Processing,* vol. 19, no. 6, pp. 1657-1663, 2010.

[26] F. Lu and J. Huang, "An improved local binary pattern operator for texture classification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 1308-1311.

[27] G. Strang, "WAVELET TRANSFORMS VERSUS FOURIER TRANSFORMS," *BULLETIN (New*

Series) OF THE AMERICAN MATHEMATICAL SOCIETY, vol. 28, no. 2, 1993.

[28] L. M. Bruce, C. H. Koger, and L. Jiang, "Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 40, no. 10, pp. 2331-2338, 2002.

[29] Y.-D. Zhang, Z. Dong, S. Wang, G. ji, and J. Yang, *Preclinical Diagnosis of Magnetic Resonance (MR) Brain Images via Discrete Wavelet Packet Transform with Tsallis Entropy and Generalized Eigenvalue Proximal Support Vector Machine (GEPSVM)*. 2015, pp. 1795-1813.

[30] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks,* vol. 13, no. 4, pp. 411-430, 2000/06/01/ 2000.

[31] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative Least Squares Regression for Multiclass Classification and Feature Selection," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 23, no. 11, pp. 1738-1754, 2012.

[32] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset," *Pattern Recognition Letters,* vol. 92, pp. 1-8, 2017/06/01/ 2017.

[33] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*. 2014, pp. 37-64.

[34] K. Hechenbichler and K. Schliep, *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*. 2004.

[35] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing,* vol. 70, no. 1, pp. 489-501, 2006/12/01/ 2006.

[36] C. W. Hsu, C. C. Chang, and C. J. Lin, *A Practical Guide to Support Vector Classification*. 2003, pp. 1396-1400.

[37] G.-B. Huang, X. Ding, and H. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing,* vol. 74, no. 1, pp. 155-163, 2010/12/01/ 2010.

[38] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," presented at the Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[39] J. Krause, A. Perer, and K. Ng, "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models," presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, California, USA, 2016. Available: https://doi.org/10.1145/2858036.2858529

[40] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence,* vol. 1, no. 5, pp. 206-215, 2019/05/01 2019.

[41] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management,* vol. 45, no. 4, pp. 427-437, 2009/07/01/ 2009.

[42] P. Mangiameli, D. West, and R. Rampal, "Model selection for medical diagnosis decision support systems," *Decision Support Systems,* vol. 36, no. 3, pp. 247-259, 2004/01/01/ 2004.

[43] L. Bottou and C.-J. Lin, "Support Vector Machine Solvers," 2007, pp. 301-320.

[44] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy," presented at the Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008, Proceedings of Machine Learning Research, 2008. Available: http://proceedings.mlr.press

[45] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4, no. 2, pp. 172-179, 1975/06/01/ 1975.

 doi:10.20944/preprints202007.0688.v1

[46] I. H. Witten, E. Frank, and M. A. Hall, "Chapter 3 - Output: Knowledge Representation," in *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, I. H. Witten, E. Frank, and M. A. Hall, Eds. Boston: Morgan Kaufmann, 2011, pp. 61-83.

[47] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," (in eng), *International journal of radiation oncology, biology, physics,* vol. 102, no. 4, pp. 1143-1158, 2018.