Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

# Performance evaluation of automated brain tumor detection systems with expert delineations and interobserver variability analysis in diseased patients on magnetic resonance imaging

Ritu Agrawal, Manisha Sharma & Bikesh Kumar Singh

**To cite this article:** Ritu Agrawal, Manisha Sharma & Bikesh Kumar Singh (2018) Performance evaluation of automated brain tumor detection systems with expert delineations and interobserver variability analysis in diseased patients on magnetic resonance imaging, Applied Artificial Intelligence, 32:7-8, 670-691, DOI: 10.1080/08839514.2018.1504500

**To link to this article:** https://doi.org/10.1080/08839514.2018.1504500

Published online: 01 Aug 2018.

APPLIED ARTIFICIAL INTELLIGENCE
2018, VOL. 32, NOS. 7–8, 670–691
https://doi.org/10.1080/08839514.2018.1504500

Taylor & Francis
Taylor & Francis Group

# Performance evaluation of automated brain tumor detection systems with expert delineations and interobserver variability analysis in diseased patients on magnetic resonance imaging

Ritu Agrawal[a], Manisha Sharma[a], and Bikesh Kumar Singh[b]

[a]Department of Electronics and Telecommunications, Bhilai Institute of Technology, Durg, Chhattisgarh, India; [b]Department of Biomedical Engineering, National Institute of Technology, Raipur, Chhattisgarh, India

**ABSTRACT**

Intervention by human expert has turned out to be essential for computerized analysis systems desiring to be approved by medical regulatory bodies. Further, to validate the performance of automated diagnosis systems, interobserver variability analysis is critically important. The purpose of this article is twofold: (i) firstly to perform interobserver variability analysis of two experienced Radiologists interpreting lesion boundary in brain magnetic resonance images; (ii) secondly, to evaluate the performance of recently proposed automated lesion segmentation model with that of the two experienced Radiologists who performed manual delineations of lesion boundary. Experiments were conducted on the database consisting of 80 real-time brain images with glioma tumor acquired using magnetic resonance imaging (MRI). Extensive statistical analysis such as the two tailed T-test, analysis of variance (ANOVA) test, Mann-Whitney U test, regression and correlation tests, etc. are performed to compare the lesions detected manually by experts and that by the automated method. Furthermore, three quantitative measures namely, dice similarity index, Jaccard coefficient, and Hausdorff distance are used to evaluate the automated lesion detection method. The experimental results show that the lesion boundaries detected by the automated method are very close to the manual delineations provided by the expert Radiologists. It is concluded that the automated systems for brain lesion detection can be utilized as a part of routine clinical practice to help the medical professionals in determining the exact location and area of lesions in brain MRI images.

## Introduction

Brain diseases such as severe meningitis, stroke, later stages of Alzheimer's illness, and brain malignancy are among the fundamental causes of death around the globe. Among these life-threatening diseases, cancer is one of the

**CONTACT** Bikesh Kumar Singh ✉ bsingh.bme@nitrr.ac.in 🖃 Department of Biomedical Engineering, National Institute of Technology, Raipur, G. E. Road Raipur, 492010.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

main causes of death during recent years. According to World Health Organization (WHO), in 2015, approximately 8.8 million deaths in the globe is due to cancer, which is approximately 15% of total deaths worldwide (World health organization cancer factsheets online). Deaths due to cancer are anticipated to increment later on, with an expected value of 11 million people in the year 2030 (World health organization cancer factsheets online). According to the statistics published by the National Brain Tumor Society, the 5-year endurance rate for the individuals with brain cancer is 34% and 36% for men and women, respectiviely (brain tumor information online). Therefore, early detection plays vital role in treatment planning of this disease (Jaya and Thanushkodi 2011). Brain tumors are predominantly categorized as benign or malignant based on their development pattern. Benign tumors develop slowly and do not proliferate to the enclosing tissues while malignant tumors are dangerous cancerous tumors which expands fast, are aggressive, and may invade nearby organs (Arakeri and Reddy 2015).

Among the different imaging modalities for detection of brain tumors, magnetic resonance imaging (MRI) is the most popular and commonly acknowledged one due to high resolution of the images obtained (El-Dahshan et al. 2014). However, manual examination of MRI images for lesion detection and quantification may turn out to be tiresome and time consuming due to the limitations posed by huge number of images, lack of expertise, and lack of manpower. Furthermore, it is subjected to observational errors and observational variabilities. The solution to this downside is the so-called computer-aided diagnosis (CAD) systems refined in the preceding few decades. CAD systems are expected to help radiologists in deriving objective evidences during the examination of medical images. Recent developments on CAD systems reveal that: (i) it can improve the diagnostic performance of radiologists; (ii) it can reduce the workload burden of medical professionals and technicians; and (iii) it can reduce diagnostic errors and improve the inter- and intraobservibility (Fujita et al. 2008; Marshkole, Singh, and Thoke 2011). However, the performance of CAD systems developed in recent years are still to be acknowledged by medical authorities because of the restrictions imposed by varying image quality, size of data, and low confidence of medical professionals on automated diagnosis systems. Thus, validation of the performance of the automated detection and diagnosis systems by expert medical professionals has turned out to be an essential research area (Singh et al. 2017).

This paper evaluates the performance of a computerized brain lesion segmentation system with that of two expert radiologists. The objective is to validate the efficacy of CAD systems for brain lesion detection by comparing its output with manual delineations provided by two expert radiologists. Additionally, we also compare the manual delineations provided by two radiologists to assess the interobserver variability. Extensive statistical and quantitative analysis is performed to study the similarities and variability between manual delineations and output of automated delineation system. Popular statistical significance analysis methods such as ANOVA test, T-test, correlation test, Mann-Whitney U test, and regression test are used for evaluation. The quantitative analysis measures such as dice similarity index, Jaccard coefficient, Hausdorff distance, segmentation accuracy, area under the curve, and image overlay plots are used to demonstrate the agreement between the automated method and the manual delineations (Chawla and Sondhi 2011; Hollander and Wolfe 1999; Jackson 2015).

## Related work

In this section, we present some relevant studies reported in context to proposed study. Santos et al. (2004) (Santos et al. 2004) proposed a lung detection method in images of x-ray computed tomography using gray level thresholding. Six radiologists were considered for interobserver analysis and among these six radiologists, two radiologists were considered for intraobserver analysis. A comparative analysis between the observers/experts manual tracing of lung delineations using different performance measures, namely, Pratt's' figure of merit, mean distance, and maximum distance was carried out. Sampat et al. 2006, performed an intra- and interobserver variability analysis with a simulated as well as real mammogram images based on Dice Similarity Coefficient (DSC) and complex wavelet structured similarity index (CW-SSIM) to identify and localize structures on the rotated and translated images. The database of 12 images with single lesions collected from the Digital Database for Screening Mammography (DDSM) were used in the analysis. Their results indicate that the intraobserver variations between two radiologists were more significant than the interobserver variations. Luijnenburg et al. (2010) (Luijnenburg et al. 2010), performed intra- and interobserver variability of biventricular function, volumes, and masses in a heterogeneous group of patients with congenital heart disease using Cardiovascular Magnetic Resonance (CMR) images. Thirty-five patients with congenital heart disease were included in the study. The performance was measured using coefficient of variation for left ventricle (LV) and right ventricle (RV). The inter- and intraobserver variability within 2.9%–6.8% and 3.9%–10.2%, respectively, was reported in their results. It was found that observer variations are higher in the intertracing category.

Nery et al. (2012) (Nery et al. 2012) reported a fully automatic approach called marker-driven watershed segmentation for detection of the pulmonary boundaries in PET images. Furthermore, a technique to delineate lung border in CT images was also reported. The results of both these techniques were also compared with manual delineations provided by several physicians using performance measures such as Pratt's measure and mean error. El-Dahshan et al. (2014) (El-Dahshan et al. 2014), proposed a model for categorization of tumors in brain MRI images. The feedback pulse coupled neural network technique was employed for detection of brain tissue followed by extraction of features using wavelet transform in discrete domain. Total of 101 MRI images of brain were used in experiments. The automated method achieved accuracy, true positive rate, and true negative rate of 99%, 92%, and 100%, respectively. The author claims that the automated method can successfully separate the healthy and pathological cases and can increase the diagnostic performance of human brain abnormality.

Arakeri et al. (2015) (Arakeri and Reddy 2015), proposed a CAD system for classification of benign and malignant brain tumors in brain MRI images using integrated classifier technique. The MRI brain dataset consisting of on 550 patients (246 females and 304 males) was obtained from the Shirdi Sai Cancer Hospital, Manipal, India. The different steps include feature extraction, feature selection and classification using artificial neural network (ANN), support vector machine (SVM), and K-Nearest neighbor (K-NN). A comparative study was carried out between performance of three radiologists and the performance of radiologists with CAD system in detecting the tumor. The performance was evaluated in terms of percentage volume overlap, Hausdorff distance, symmetric mean absolute surface distance, accuracy, true positive rate, true negative rate, and area under receiver operating characteristics (AUC). The automated CAD system achieved a classification accuracy of 99.09%, sensitivity of 100%, specificity of 98.21%, and AUC of 0.991.The AUC achieved by four radiologists were 0.995, 0.995, 0.786, and 0.731, respectively.

Büyükdereli et al. (2016) (Büyükdereli and Güler 2016) did a variability analysis of two expert nuclear medicine physicians in evaluating the metabolic and morphologic prominence of lung malignancies. The study was conducted on positron emission tomography/computed tomography (PET/CT) images which provide valuable information about differential finding, staging, and therapy reaction. The dataset used consisted of 97 patients taken from the Nuclear medicine department of State Hospital, Nigde Turkey. The performance was evaluated in terms of the coefficient of correlation, regression test, and Bland-Atman plot. The coefficient of correlation value of 0.96 and 0.96 was obtained for inter- and intraobserver variability respectively. Hsieh et al. (2017) (Hsieh et al. 2017), proposed a CAD system based on machine learning scheme for classification of low and high grade Gliomas (a malignant tumor). The MRI datasets consisting of 34 Glioblastomas (high grade tumor) and 71 Gliomas (low-grade tumor) obtained from the cancer imaging archive (TCIA) open source were used in experiments. Histogram based features were used in the analysis. The performance of the automated approach in grading Gliomas in MR images was compared with the radiologists' performance using accuracy, true positive rate, true negative rate, and AUC. The corresponding values achieved were 87%, 90%, 79%, and 0.89, respectively.

Having reviewed the aforementioned studies, we conclude that expert variability investigation is conducted on some imaging modalities such as mammography for detection of lesions, quantifying congenital heart disease in CMR images, detecting malignancies in lung PET/CT images, etc. Most of the studies concerning computerized brain MRI image analysis concentrate on classification of lesions in brain MRI images. However, very limited studies are conducted on inter- and intraobserver analysis of brain lesion detection in MRI images. Further, comparing performance of computer-aided automated lesion segmentation in brain MRI images with manual delineations provided by radiologists is also little reported. This article aims to address these issues. The contributions of this paper are highlighted in the next section.

### *Contributions and organisation of the manuscript*

The contributions of this article are now encapsulated as follows:

(a) We implement and evaluate the performance of automated lesion detection systems in brain MRI images with manual delineations provided by two expert radiologists.
(b) We investigate the interobserver variability between manual delineations provided by two expert radiologists in detecting lesions in brain MRI images.
(c) The automated and manual delineations are evaluated and quantified using several statistical and quantitative analysis methods.

The remaining part of this article is structured as follows. Section 2 elucidates materials and methods employed in this work. Results and discussions are presented in section 3 followed by conclusions and future scopes in section 4. Table 1 illustrates the abbreviations utilized in this paper.

## Material and methods

### *Data acquisition*

The dataset used in this study comprised of axial, T1 and T2 weighted brain MRI images obtained from 80 subjects (48 males and 32 females) between the ages of 20 and 65 years. The images were collected from local scan center Imaging Point, Nagpur, India and some online resources such as Brain Web images (BrainWeb 2017) and BRATS MRI images (Menze et al. 2015). The images collected from local scan center imaging point and online Brain Web images

**Table 1.** Abbreviations and definition used in this study.

| Abbreviations | Definition | Equation | Remarks |
|---|---|---|---|
| Exp-1 | Expert/Observer 1 | - | - |
| Exp-2 | Expert/Observer 2 | - | - |
| ANOVA | Analysis of Variance | - | - |
| DSC | Dice Similarity Coefficient | $\mathrm{DSC} = 2\frac{\lvert \mathrm{A} \cap^{\mathrm{B}} \rvert}{\lvert \mathrm{A} \rvert + \lvert \mathrm{B} \rvert}$ | The values of DSC and JI lies between “0” and “1.” “0” represents no-overlapping between the two images and “1” represents the perfect overlapping. |
| JI | Jaccard Index | $JI = \left\lvert \frac{A \cap^{B}}{A \cup^{B}} \right\rvert$ | |
| HD | Hausdorff Distance | $HD = \max(h(A,B), h(B,A))$ | - |
| TP | True positive | $TP = \frac{\eta_{tp}}{\eta}$ | $\eta_{tp}$is the total number of pixels in tumor region while $\eta$is the total number of pixels in the image. |
| TN | True Negative | $TN = \frac{\eta_{tn}}{\eta}$ | $\eta_{tn}$is the total number of pixels in the nontumor region |
| FP | False positive | $TN = \frac{\eta_{fp}}{\eta}$ | $\eta_{fp}$ is the number of pixels corresponding to nontumor region wrongly detected in tumor region. |
| FN | False Negative | $TN = \frac{\eta_{fn}}{\eta}$ | $\eta_{fn}$is the number of pixels corresponding to tumor region wrongly detected in nontumor region. |
| $\mathrm{S_p}$ | Specificity | $S_p = \frac{TN}{TN+FP}$ | It is the percentage of correctly detected pixels corresponding to nontumor region. Higher values of $\mathrm{S_p}$are desirable. |
| $\mathrm{S_e}$ | Sensitivity | $S_e = \frac{TP}{TP+FN}$ | It is the percentage of correctly detected pixels corresponding to tumor region Higher values of $\mathrm{S_e}$ are desirable. |
| SA | Segmentation Accuracy | $SA = \frac{TP+TN}{TP+FN+TN+FP}$ | It is the percentage of correctly detected pixels. Higher values of SA are desirable |
| AUC | Area Under the receiver operating Curve/ Characteristics | $AUC = \frac{1}{2}(S_p + S_e)$ | It is a common measure of sensitivity and specificity. Higher values of AUC are desirable. |

“A” represents the segmented tumor output of the automated method; “B” represents manually segmented tumor output by observers.

(BrainWeb 2017), were acquired using 1.5 Tesla MRI scanner. Furthermore, BRATS MRI images (Menze et al. 2015) were acquired using (1.5–3) Tesla MRI scanner. The local scan center images were recorded during a time period of December 2016 to May 2017. The acquired images were resized to size 256 × 256 pixels to improve computational efficiency of the automated system. All the 80 patients were diagnosed with malignant tumor (gliomas tumor). Figure 1, shows some of the images from the data set with lesion area pointed by arrow.

## Data preprocessing

Before automated segmentation of brain MRI images, preprocessing is usually required to improve the quality of images. This step includes impulse noise removal and skull stripping to facilitate accurate lesion

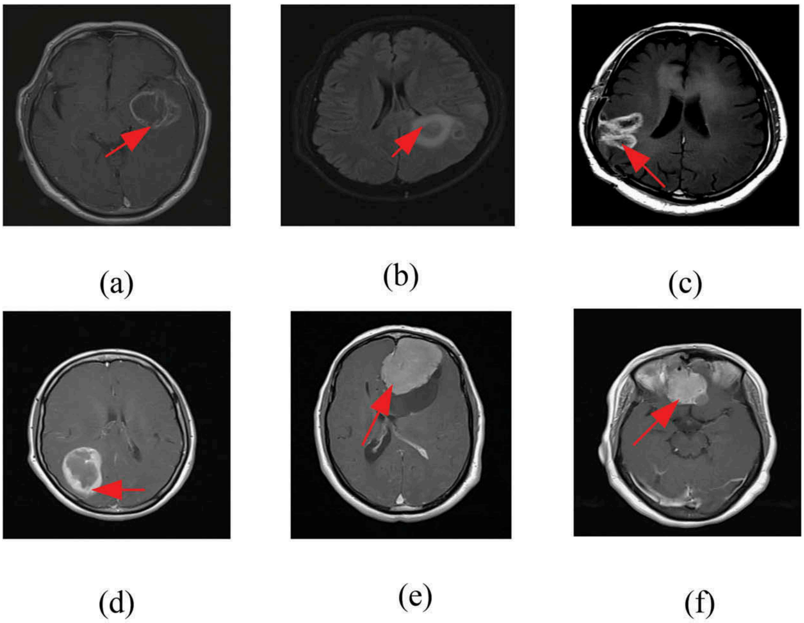


**Figure 1.** Some examples of MRI images consisting of gliomas tumor from dataset. (a and b) T1 weighted MRI image(c–f) T2 weighted MRI image.

detection. For noise removal, a noise removal filter based on biorthogonal wavelet transform is used in this study due to its high edge preserving characteristics (Prakash, Khare, and Khare 2013). The noise removal process consists of three steps: (i) the brain MRI images are initially decomposed into four bands, namely, LL, LH, HL, and HH, by using discrete wavelet transform (DWT). The mother wavelet used in this study is biorthogonal bior 1.3; (ii) hard thresholding is then applied to eliminate noisy wavelet coefficients; and (iii) the denoised image is then reconstructed using inverse DWT.

Skull stripping involves removal of non-cerebral matters such as the cranium, scalp, etc., since these tissues usually does not contain any valuable information. To perform this operation, a segmentation technique based on threshold is utilized in the experiments. It includes the following steps: (i) initially, the acquired images are transformed into grayscale images. Then by using appropriate threshold, the grayscale images is transformed into binary images; (ii) the resulting image is further processed by preserving the preset threshold level and discarding other pixel values to generate a binary mask; (ii) the image with stripped skull is then determined by multiplying the binary mask with the input image. The details of experiments and results can be found in (Agrawal, Sharma, and Singh 2017).

## Automated lesion detection in brain MRI images

This section explains the recently proposed automated method for lesion detection in brain MRI images (Agrawal, Sharma, and Singh 2017). The performance of the automated lesion detection method is compared with that of manual delineations provided by two expert radiologists for validating the efficacy of automated method. The framework of the automated lesion detection system is shown in Figure 2. It comprises of five stages, namely: (i) image acquisition and preprocessing using biorthogonal wavelet transform

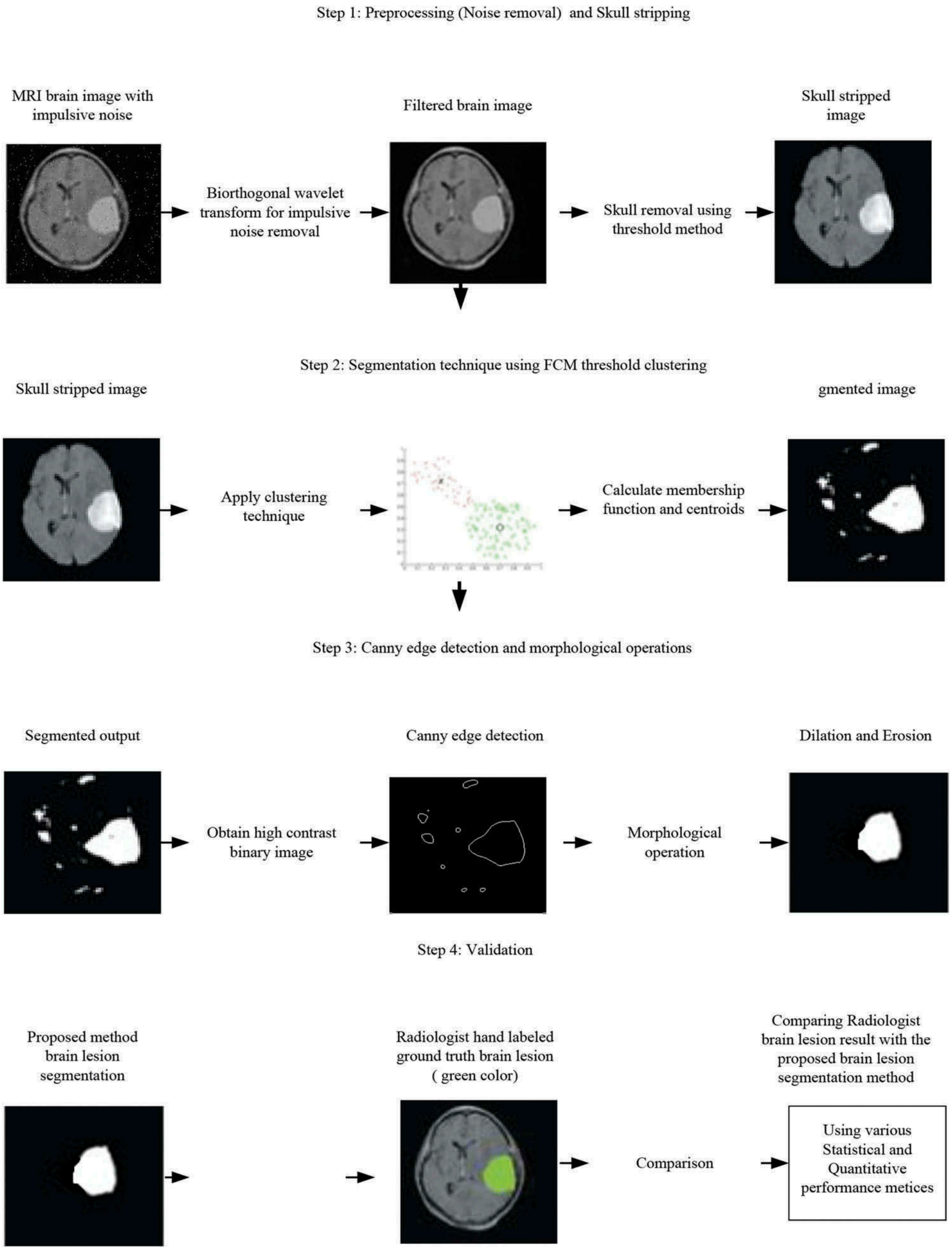

Step 1: Preprocessing (Noise removal) and Skull stripping
MRI brain image with impulsive noise
Biorthogonal wavelet transform for impulsive noise removal
Filtered brain image
Skull removal using threshold method
Skull stripped image
Step 2: Segmentation technique using FCM threshold clustering
Skull stripped image
Apply clustering technique
Calculate membership function and centroids
gmented image
Step 3: Canny edge detection and morphological operations
Segmented output
Obtain high contrast binary image
Canny edge detection
Morphological operation
Dilation and Erosion
Step 4: Validation
Proposed method brain lesion segmentation
Radiologist hand labeled ground truth brain lesion ( green color)
Comparison
Comparing Radiologist brain lesion result with the proposed brain lesion segmentation method
Using various Statistical and Quantitative performance metices

Figure 2. Flow diagram of automated method of brain lesion delineation.

for noise removal, (ii) skull removal using threshold based approach; (iii) clustering by fuzzy c-means for initial segmentation; (iv) edge detection using Canny edge detector followed by morphological operations for extracting the final region of interest (ROI); and (v) performance evaluation using various statistical and quantitative measures. Detailed description of the experiments and results can be found in (Agrawal, Sharma, and Singh 2017).

## Manual delineation

In order to evaluate the performance of the automated delineation system, manual delineations provided by two observers were utilized in this study. Both observers are expert/trained radiologists with an experience of thirty-five years and twenty years, respectively, in analyzing brain MRI images. The tracing of lesions in MRI images were done manually and independently using MATLAB® software. Figure 3 shows the manual tracing of the brain tumor border by two observers.

## Automated system for interobserver analysis

To perform interobserver analysis, various techniques of statistical significance analysis and quantitative analysis are used. The objectives of the interobserver analysis are:

- To compare the performance of both expert observers in delineating the lesion boundary in MRI images.
- To compare the performance of the automated lesion detection system with that of expert observers.

In order to compare the results of automated delineation with manual delineation, the area of detected tumor in brain MRI images is calculated in mm$^2$ using following equation:

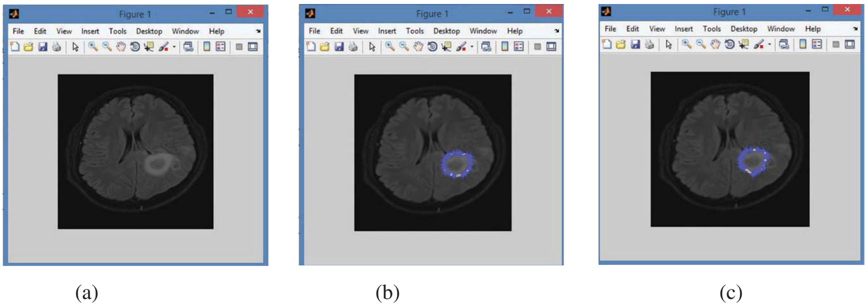


Figure 3. Manual tracing of brain lesion using MATLAB software. (a) Original real MRI image. (b) Manual tracing by Exp-1. (c) Manual tracing by Exp-2.

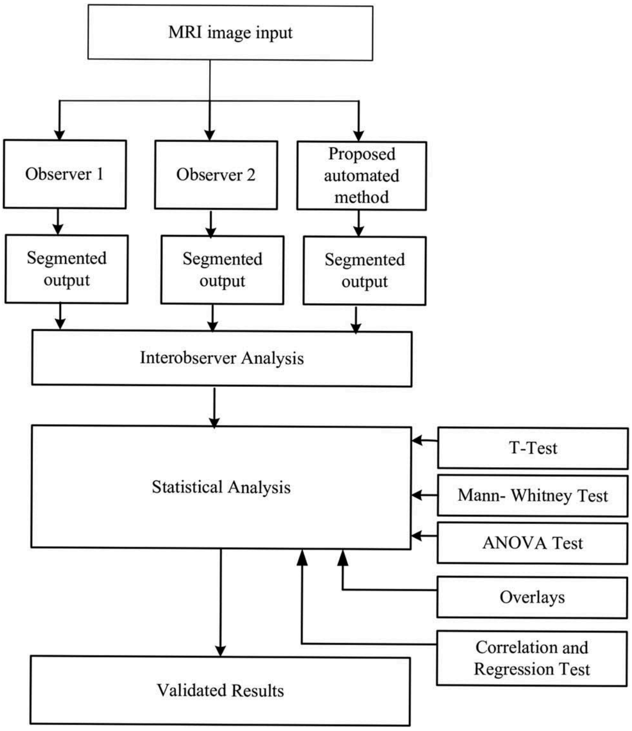


**Figure 4.** Overview of automated interobserver variability system.

$$\text{Area}(\text{mm})^2 = \sqrt{x} * 0.264 \tag{1}$$

where $x$ represents number of white pixels.

Figure 4 shows the block diagram of automated interobserver variability system. The input to this system is brain MRI image containing malignant lesion. The input MRI image is analyzed by both the observers for tracing tumor boundary manually. The same input image is processed through automated delineation system. The output of the two observers and the automated system are then compared using statistical and quantitative methods. The different evaluation measures used are:

- The descriptive statistics such as mean, median, variance, standard deviation, standard mean error, maximum, and minimum value.
- Statistical significance analysis using ANOVA test, T-test, Mann-Whitney U test, coefficient of correlation, and regression test.
- Performance measures such as DSC, JI, HD, SA, sensitivity, specificity, and AUC.

## Results and discussions

Computer aided automated detection systems have come out as an alternative tool for locating abnormal tissues in medical images. The main purpose of such systems is to assist the radiologists in their routine clinical practice by providing more objective evidences. However, the performance of the current automated lesion detection systems do not satisfy the needs of real time application due to low confidence of radiologists on CAD systems and unseemly validation of segmentation algorithms. This article compares the performance of a recently proposed automated lesion detection system with manual delineations provided by expert radiologists. We hypothesize that automated lesion detection systems can be employed in routine clinical practice only if their performance is close to radiologist's performance. This is also important to gain trust of radiologists in such automated systems.

This section presents the results of variability analysis within observers and between the results of automated and manual method. The statistical significance analysis is carried out using SPSS® (Statistical package for the Social Sciences) software and the quantitative analysis is performed using MATLAB® R2010a, software.

### Descriptive analysis

Table 2 presents the results of descriptive statistical analysis of MRI brain tumor extraction between the two observers and the automated method to illustrate the variations between them. It is found that the mean of tumor area delineated by both the observers and the automated method are very close, that is, 97.703, 97.708, and 97.715 mm$^2$. The standard deviation and the mean of standard error are found to be 41.737 and 13.198, respectively, for Exp-1, 41.734 and 13.197, respectively, for Exp-2 while 41.732 and 13.196, respectively, for automated method. It is thus concluded that the descriptive statistics of the two observers and the automated method for brain lesion extraction are very close to each other resulting in lower variability between them. However, this variability may increase if opinion of more number of observers are included in the study.

Table 2. Descriptive statistics of lesions detected by two experts and by automated method.

| | Tumor area of MRI brain (100 mm$^2$) | | | | | | |
|---|---|---|---|---|---|---|---|
| Method | Mean | Median | Variance | Standard deviation | Standard error mean | Minimum | Maximum |
| Exp-1 | 0.977 | 0.791 | 1.742 | 0.417 | 0.131 | 0.784 | 2.140 |
| Exp-2 | 0.978 | 0.791 | 1.741 | 0.417 | 0.131 | 0.784 | 2.140 |
| Automated method | 0.977 | 0.792 | 1.741 | 0.417 | 0.131 | 0.784 | 2.140 |

## Evaluation using statistical significance analysis

### ANOVA test between experts and the automated method

Analysis of Variance (ANOVA) test is carried out between the manual delineations provided by the two observers and that by automated method considering area of detected brain lesion as feature. The results are depicted in Table 3. It is found that the *P* value obtained is 0.911 indicating that the lesion segmented by the automated method is very similar to the manual delineations provided by the two observers in terms of area of the detected tumor.

*3.2.2 T-test and Mann-Whitney U test between observers and the automated method and its interpretation*

This section present the results of the two tailed *T*-test and Mann-Whitney U test when applied to analyze the variability between the manual delineations provided by two observers and the detected lesion by the automated method. The variability analysis is carried out between following three categories: (i) manual delineations provided by the observer 1 (Exp-1) and 2 (Exp-2); (ii) the manual delineation provided by observer 1 and the lesion segmented by the automated method; and (iii) the manual delineation provided by observer 2 and the lesion segmented by the automated method. The results are shown in Table 4. It is observed that the P-value obtained for two-tailed T test between Exp-1 and Exp-2, Exp-1 and the automated method, and between Exp-2 and the automated method is found to be 0.887, 1, and 0.877, respectively. The *P* value obtained in all the three cases indicates that the segmentation results of the two observers and that of the automated method for brain lesion extraction are very close to each other. However, the performance of automated lesion detection system is more close to the manual delineations provided by Exp-1, that is, the more experienced radiologists resulting in higher *P* value of 1.

**Table 3.** The ANOVA calculation for brain tumor extraction using two observers and the automated method.

| Category | Sum of squares | Degree of freedom (DF) | Mean square | F | *P* |
|---|---|---|---|---|---|
| Between groups | 18.000 | 28 | 0.643 | 0.321 | 0.911 |
| Within groups | 2.000 | 1 | 2.000 | | |
| Total | 20.000 | 29 | | | |

**Table 4.** *T*-test and Mann-Whitney test results in MRI brain of interobserver variability.

| | *T*-test | Mann-Whitney test | |
|---|---|---|---|
| Source | *P* value | Z statistic value | *P* value |
| Exp-1 vs. Exp-2 | 0.887 | −0.151 | 0.880 |
| Exp-1 vs. automated | 1.000 | −0.038 | 0.970 |
| Exp-2 vs. automated | 0.877 | −0.302 | 0.862 |

In case of Mann-Whitney U test conducted for the three categories as discussed earlier, it is observed from Table 4 that the Z statistic value obtained for Exp-1 and Exp-2, Exp-1 and automated method, and Exp-2 and automated method is found to be −0.151, −0.038, and −0.302, respectively. Further, the *P* value for these three categories are found to be is 0.880, 0.970, and 0.862, respectively. From these values, it is concluded that the segmented lesion in all the three cases are very similar to each other, however, as in previous case the level of agreement between the Exp-1 and the automated method is more promising. This may be due to the fact that the Exp-1 is more experienced than the Exp-2.

*3.2.3 Coefficient of correlation and regression test between the experts and the automated method and its interpretation*

The results pertaining to analysis of coefficient of correlation and regression for three categories namely, Exp-1 and Exp-2, Exp-1 and the automated method and Exp-2 and the automated method are shown in Table 5. The value of coefficient of correlation are found to be 0.987, 1, and 0.984, respectively, for the three categories. The coefficient of correlation between segmented lesions within all the three categories is high indicating high similarity between performance of manual and automated approach. It is further observed that the performance of automated lesion detection system is more closer to that of Exp-1, that is, the more experienced radiologist as in previous tests.

The regression analysis between segmented lesions corresponding to three categories yield R-squared value of 0.870, 1, and 0.812, respectively. This indicates strong association between lesions segmented by automated method and that by Exp-1, that is, the more experienced radiologist. Furthermore, a small interobserver variability is found between Exp-1 and Exp-2 resulting in R-squared value of 0.87.

## Dice similarity coefficient and Jaccard index between experts and the automated method

To further illustrate the utility of automated lesion detection systems, two popular similarity measures, namely, JI and DSC are used (Prabha and Kumar 2016). The measures JI and DSC are calculated for three different categories, that is, between the lesions detected by the Exp-1 and Exp-2, Exp-

**Table 5.** Correlation coefficient and regression test results in MRI brain of interobserver variability.

| | Coefficient of correlation analysis | Regression analysis |
|---|---|---|
| Source | Correlation value | R-squared value |
| Exp-1 vs. Exp-2 | 0.987 | 0.870 |
| Exp-1 vs. automated | 1.00 | 1.000 |
| Exp-2 vs. automated | 0.984 | 0.812 |

Table 6. Similarity index measures between experts and the automated method.

| | Similarity measure | | | | | |
|---|---|---|---|---|---|---|
| | Jaccard index | | | Dice similarity coefficient | | |
| MRI images | Exp-1 and automated | Exp-2 and automated | Exp-1 and Exp-2 | Exp-1 and automated | Exp-2 and automated | Exp-1 and Exp-2 |
| #1 | 0.9485 | 0.9004 | 0.9193 | 0.9735 | 0.9475 | 0.9579 |
| #2 | 0.8695 | 0.8606 | 0.8993 | 0.9301 | 0.9251 | 0.9469 |
| #3 | 1 | 0.9902 | 0.9707 | 1 | 0.9951 | 0.9851 |
| #4 | 0.9297 | 0.8361 | 0.9558 | 0.9635 | 0.9107 | 0.9774 |
| #5 | 0.9355 | 0.9023 | 0.8578 | 0.9667 | 0.9486 | 0.9234 |
| #6 | 0.9914 | 0.9695 | 0.9262 | 0.9956 | 0.9845 | 0.9616 |
| #7 | 0.9341 | 0.9205 | 0.9448 | 0.9659 | 0.9586 | 0.9716 |
| #8 | 0.9332 | 0.8964 | 0.8944 | 0.9654 | 0.9453 | 0.9442 |
| #9 | 0.8710 | 0.8735 | 0.914 | 0.9310 | 0.9324 | 0.955 |
| #10 | 0.9393 | 0.9233 | 0.8847 | 0.9687 | 0.9601 | 0.9388 |
| Mean | 0.9352 | 0.9072 | 0.9167 | 0.9660 | 0.9507 | 0.9561 |

1 and the automated system and Exp-2 and the automated system as discussed in previous sections. The JI and DSC values obtained for ten different MRI brain images are shown in Table 6. The results show that there is a high similarity between the lesions detected by Exp-1 and automated method achieving highest JI and DSC values of 0.9352 and 0.9660, respectively. Further, the lesion detected by automated method is more closer to that detected by Exp-1 as compared to that by Exp-2. These results are also in agreement with the other tests conducted in this study.

## Performance of the automated method against two experts using Hausdorff distance

Another popular similarity measure used to compare the performance of the automated lesion detection system with that of the manual system is Hausdorff distance. The detailed description of this performance measure can be found in (Beauchemin, Thomson, and Edwards 1998; Huttenlocher, Klanderman, and Rucklidge 1993; Saba et al. 2016) and the corresponding results are shown in Table 7. It is found that the lowest mean value of HD, that is, 2.753 is achieved between lesions segmented by Exp-1 and the automated method while the highest mean value of HD, that is, 2.830 is obtained between the lesions detected by Exp-2 and the automated method. The results thus indicate strong association between the lesions segmented by manual and automated approaches.

*3.2.6 Performance evaluation of the automated method against two experts using sensitivity, specificity, segmentation accuracy and area under the curve*

The measures such as sensitivity, specificity, segmentation accuracy (SA), and area under the curve (AUC) are extremely important to access the performance of computer aided detection and diagnosis systems (Van Erkel and Pattynama 1998). These measures are usually obtained from the confusion matrix defined in terms of TP, TN, FP, and FN. The formula used for

**Table 7.** Hausdorff distance between experts and the automated method.

| | Hausdorff distance (mm) | | |
|---|---|---|---|
| MRI images | Exp-1 and automated | Exp-2 and automated | Exp-1 and Exp-2 |
| #1 | 2.449 | 2.645 | 2.645 |
| #2 | 3.873 | 3.873 | 3.761 |
| #3 | 0 | 0.023 | 0.036 |
| #4 | 4.472 | 4.502 | 3.986 |
| #5 | 3.873 | 3.982 | 3.991 |
| #6 | 1.732 | 1.753 | 1.80 |
| #7 | 3.605 | 3.605 | 3.529 |
| #8 | 2.236 | 2.605 | 2.765 |
| #9 | 3.162 | 3.208 | 3.321 |
| #10 | 2.136 | 2.108 | 2.328 |
| Mean | 2.753 | 2.830 | 2.816 |

the calculation of sensitivity, specificity, SA, and AUC are defined in Table 1 and the results obtained for three categories are shown in Table 8 (a-d). It is observed that the mean values of specificity, sensitivity, SA, and AUC are found to be 0.998, 0.957, 0.996, and 0.977, respectively, for the first category, that is, between Exp-1 and Exp-2. On the other hand, these values for category two (Exp-1 and automated) and three (Exp-2 and automated) are found to be 0.995, 0.996, 0.995, 0.996 and 0.999, 0.899, 0.994, 0.94, respectively. The result shows that for all the three cases SA achieved is above 99%

**Table 8(a).** Specificity analysis between experts and automated method.

| MRI images | Exp-1 and Exp-2 | Exp-1 and automated | Exp-2 and automated |
|---|---|---|---|
| #1 | 0.999 | 0.998 | 0.999 |
| #2 | 0.998 | 0.996 | 0.999 |
| #3 | 1.00 | 1.00 | 0.999 |
| #4 | 0.998 | 0.990 | 0.999 |
| #5 | 1.00 | 0.997 | 0.999 |
| #6 | 0.999 | 0.994 | 1.00 |
| #7 | 0.998 | 0.995 | 0.999 |
| #8 | 0.997 | 0.993 | 0.999 |
| #9 | 0.993 | 0.996 | 0.995 |
| #70 | 0.997 | 0.995 | 0.999 |
| Mean | 0.998 | 0.995 | 0.999 |

**Table 8(b).** Sensitivity analysis between experts and automated method.

| MRI images | Exp-1 and Exp-2 | Exp-1 and automated | Exp-2 and automated |
|---|---|---|---|
| #1 | 0.949 | 1.00 | 0.900 |
| #2 | 0.907 | 1.00 | 0.866 |
| #3 | 1.00 | 0.970 | 1.00 |
| #4 | 0.943 | 1.00 | 0.845 |
| #5 | 0.935 | 0.991 | 0.905 |
| #6 | 1.00 | 1.00 | 0.869 |
| #7 | 0.955 | 1.00 | 0.921 |
| #8 | 1.00 | 1.00 | 0.897 |
| #9 | 0.885 | 1.00 | 0.854 |
| #10 | 1.00 | 1.00 | 0.932 |
| Mean | 0.957 | 0.996 | 0.899 |

**Table 8(c).** Segmentation accuracy between experts and automated method.

| MRI images | Exp-1 and Exp-2 | Exp-1 and automated | Exp-2 and automated |
|---|---|---|---|
| #1 | 0.998 | 0.998 | 0.997 |
| #2 | 0.994 | 0.996 | 0.994 |
| #3 | 1.00 | 0.999 | 0.999 |
| #4 | 0.994 | 0.991 | 0.986 |
| #5 | 0.995 | 0.997 | 0.993 |
| #6 | 0.999 | 0.994 | 0.994 |
| #7 | 0.995 | 0.996 | 0.994 |
| #8 | 0.997 | 0.993 | .995 |
| #9 | 0.989 | 0.996 | 0.989 |
| #10 | 0.998 | 0.996 | 0.997 |
| Mean | 0.996 | 0.995 | 0.994 |

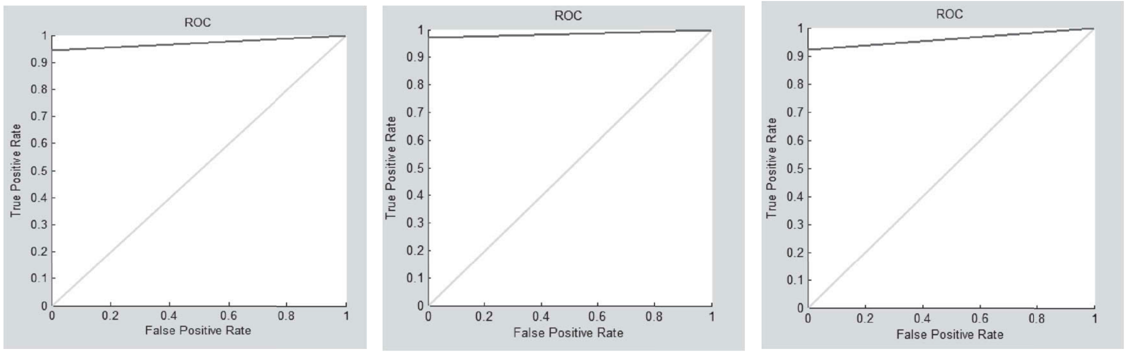
**Table 8(d).** Area under receiver operating curve analysis between experts and automated method.

| MRI images | Exp-1 and Exp-2 | Exp-1 and automated | Exp-2 and automated |
|---|---|---|---|
| #1 | 0.974 | 0.999 | 0.950 |
| #2 | 0.952 | 0.998 | 0.933 |
| #3 | 1.00 | 0.985 | 0.999 |
| #4 | 0.971 | 0.995 | 0.922 |
| #5 | 0.967 | 0.994 | 0.952 |
| #6 | 0.999 | 0.997 | 0.934 |
| #7 | 0.976 | 0.997 | 0.960 |
| #8 | 0.998 | 0.996 | 0.948 |
| #9 | 0.939 | 0.998 | 0.925 |
| #10 | 0.998 | 0.997 | 0.965 |
| Mean | 0.977 | 0.996 | 0.949 |

indicating high similarity between lesions detected by automated method and the manual delineations provided by the two experts. Figure 5 shows the ROC curve of image 2 obtained for three categories. It is found that the area under ROC for all three categories is above 94% indicating strong association between the results obtained by the automated method, Exp-1 and Exp-2.

*3.2.7 Performance evaluation of lesions detection systems using percentage overlap of segmented area*

This section evaluates the automated and manual lesion detection systems by using percentage overlap similarity between the segmented lesions. The lesions detected using different approaches were overlaid to determine the percentage overlap. The results obtained for three categories are shown in Tables 9 (a and b) and 10. The lesions detected by the two methods in a particular category are differentiated using white, green, and red color. It is observed from Table 9 that the green and white regions corresponding to detected lesion shows maximum overlapping, indicating high similarity. From Table 10, it is found that the mean percentage overlap for three categories are found to be 97%, 94.29%, and 95.90%, respectively. Hence, it is concluded that the automated method can successfully segment the lesion area with a high degree of percentage overlapping.



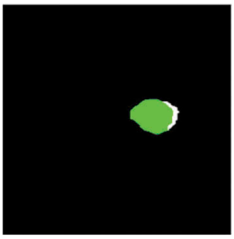
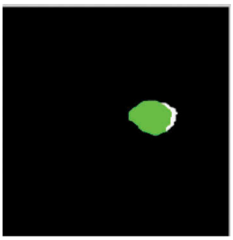
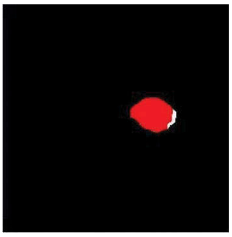
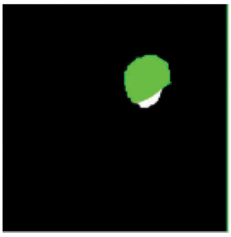
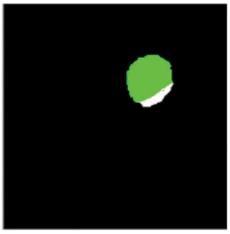
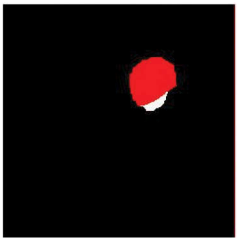
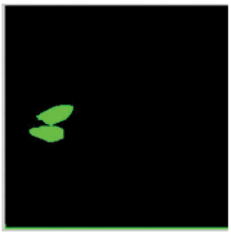
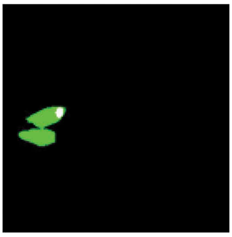
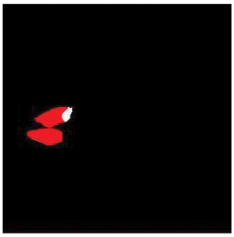
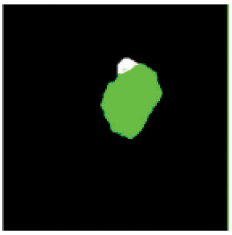
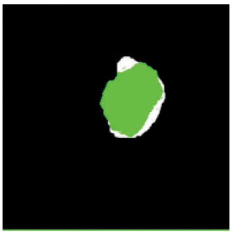
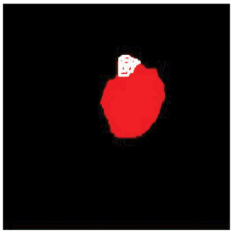
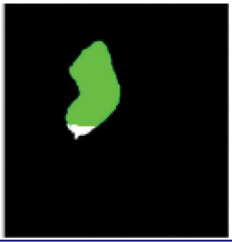
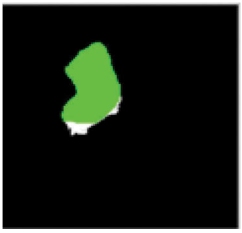
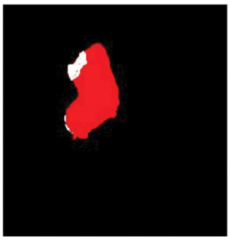
Obs-1 and Obs-2

Obs-1 and automated method

Obs-2 and automated method

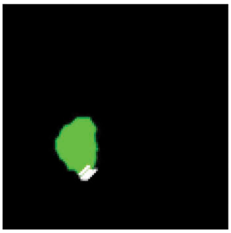
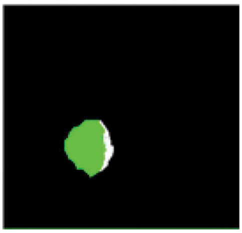
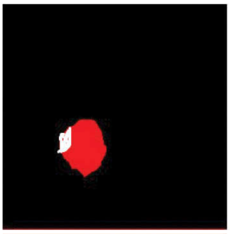
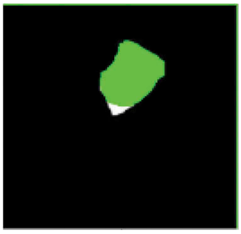
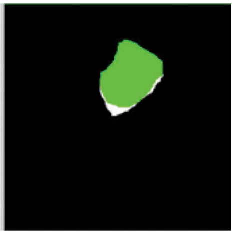
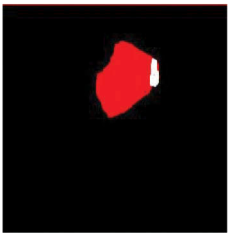
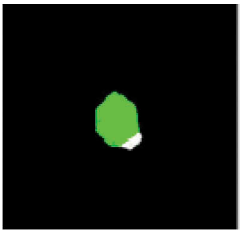
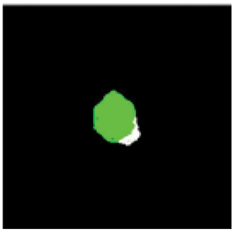
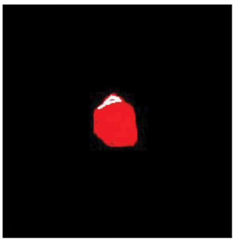
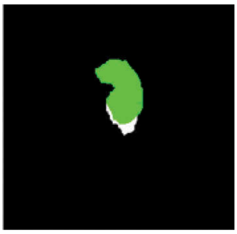
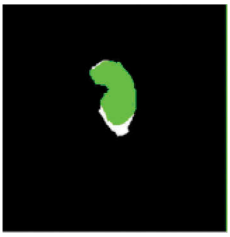
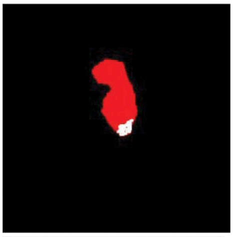
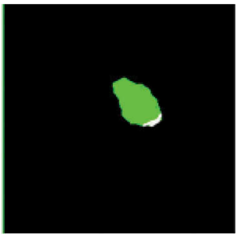
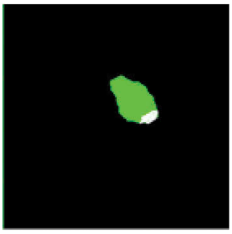
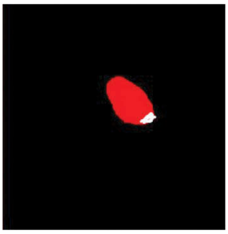
**Figure 5.** ROC curve of MRI brain for three categories.

**Table 9(a).** Overlay of detected brain lesions for ten MRI images (image 1–5).

| Real MRI brain images | Obs-1 (white) and automated method (green) | Obs-2 (white) and automated method (green) | Obs-1 (white) and Obs-2 (red) |
|---|---|---|---|
| Image 1 |  |  |  |
| Image 2 |  |  |  |
| Image 3 |  |  |  |
| Image 4 |  |  |  |
| Image 5 |  |  |  |

The important motivation behind this investigation is to explore the variability in the performance of brain MRI segmentation using manual and automated approaches. From the result section, it is observed that in terms of percentage overlay the automated segmentation method shows maximum overlapping boundaries (green color) against the manual tracing of two experts boundaries (white color) (refer table 9). The

Table 9(b). Overlay of detected brain lesions for ten MRI images (image 6–10).

| Real MRI brain images | Obs-1 (white) and automated method (green) | Obs-2 (white) and automated method (green) | Obs-1 (white) and Obs-2 (red) |
|---|---|---|---|
| Image 6 |  |  |  |
| Image 7 |  |  |  |
| Image 8 |  |  |  |
| Image 9 |  |  |  |
| Image 10 |  |  |  |

general statistics from Table 2 illustrates that the Exp-1, Exp-2 and the automated method results are in agreement with each other.

Overall, the results indicate that the automated system can successfully detect the brain lesion with an adequate accuracy and level of agreement when compared to manual delineations provided by the two experts. Involvement of experts has become more or less an important requisite for an automated diagnosis system wishing to be recognized by

Table 10. Percentage overlap of segmented lesion area in ten brain MRI images.

| | Percentage overlay (%) | | |
|---|---|---|---|
| MRI Images | Exp-1 and automated | Exp-2 and automated | Exp-1 and Exp-2 |
| Image 1 | 95.90 | 92.09 | 97.9 |
| Image 2 | 93.80 | 91.59 | 93.8 |
| Image 3 | 100 | 99.56 | 97.06 |
| Image 4 | 94.33 | 90.55 | 96.8 |
| Image 5 | 98.55 | 93.54 | 93.51 |
| Iamge 6 | 99.50 | 96.94 | 95.32 |
| Image 7 | 95.50 | 92.18 | 98.56 |
| Image 8 | 96.58 | 94.90 | 94.83 |
| Image 9 | 97.51 | 95.41 | 96.80 |
| Image 10 | 98.35 | 96.22 | 94.50 |
| Mean | 97.00 | 94.29 | 95.90 |

clinical regulatory agencies. Further, building up of confidence on information or communication technology is also an vital motivation for integrating radiologists with CAD systems. Development of future CAD systems requires association of experts from both domains, that is, medicine and technology. In brain MRI based CAD systems, experts from these domains can be implicated in various phases of CAD implementation, testing, and validation. Developing approaches to improve interaction between experts is also important for clinical approval and trust of such systems.

## Conclusions and future scopes

This study investigated an interobserver variability analysis between manually traced brain tumor border by two expert radiologist's/observers and further compared their performance with a recently proposed automated tumor detection system. Some popular statistical analysis techniques such as ANOVA test, two-tailed *T*-test, Mann-Whitney U test, correlation test, and R-squared regression test were used to compare the performance of manual and automated approach. Further, the performance measures, namely, JI, DSC, and Hausdorff distance were also used to determine the similarity between output of observers and automated method. The results indicate that: (i) the performance of automated lesion detection system is in agreement with the lesions detected manually by two experts; (ii) the results obtained using automated system are more closer to the manual delineations provided by expert 1, that is, the more experienced radiologist. It is thus concluded that computerized automated systems can be used in clinical practice for detection of brain lesions in MRI images. Such systems can assist the medical professionals in deriving objective evidences in support of diagnostic results. In future, comparative studies between radiologists performance with CAD systems based on different imaging modalities can be conducted. More number of experts can included in the study. Performance of CAD systems with less experienced, intermediate experienced, and more experienced radiologists is also looked out as future scope. Intraobserver analysis can also be conducted in future.

## References

Agrawal, R., M. Sharma, and B. K. Singh. 2017. Segmentation of brain tumor based on clustering technique: Performance analysis. *Journal of Intelligent System, DE GRUYTER* 1–16. doi: https://doi.org/10.1515/jisys-2017-0027

Arakeri, M. P., and G. R. M. Reddy. 2015. Computer-aided diagnosis system for tissue characterization of brain tumor on magnetic resonance images. *Signal, Image and Video Processing* 9 (2):409–25. doi:10.1007/s11760-013-0456-z.

Beauchemin, M., K. P. B. Thomson, and G. Edwards. 1998. On the Hausdorff distance used for the evaluation of segmentation results. *Canadian Journal of Remote Sensing* 24 (1):3–8. doi:10.1080/07038992.1998.10874685.

Brain tumor information, National Brain Tumor Society.[online]: http://www.braintumor.org/brain-tumor-information.

BrainWeb, *Simulated Brain Database*, Available at: http://brainweb.bic.mni.mcgill.ca/brainweb/, last accessed on 2 January, 2017.

Büyükdereli, G., and M. Güler. 2016. Inter-observer and Intra-observer Variability among Measurements of FDG PET/CT Parameters in Pulmonary Tumors. *Balkan Medical Journal* 33 (3):308–15. doi:10.5152/balkanmedj.2016.140530.

Chawla, D., and N. Sondhi. 2011. *Research methodology concepts and cases*. New Delhi: Vikas publishing house private limited.

El-Dahshan, E. S. A., H. M. Mohsen, K. Revett, and A. B. M. Salem. 2014. Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert Systems with Applications* 41 (11):5526–45. doi:10.1016/j.eswa.2014.01.021.

Fujita, H., Y. Uchiyama, T. Nakagawa, D. Fukuoka, Y. Hatanaka, T. Hara, … X. Zhou. 2008. Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs. *Computer Methods and Programs in Biomedicine* 92 (3):238–48. doi:10.1016/j.cmpb.2008.04.003.

Hollander, M., and D. A. Wolfe. 1999. *Nonparametric statistical methods*. New York: Wiley.

Hsieh, K. L. C., R. J. Tsai, Y. C. Teng, and C. M. Lo. 2017. Effect of a computer-aided diagnosis system on radiologists' performance in grading gliomas with MRI. *PloS One* 12 (2):14. pages. doi:10.1371/journal.pone.0171342.

Huttenlocher, D. P., G. A. Klanderman, and W. J. Rucklidge. 1993. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (9):850–63. doi:10.1109/34.232073.

Jackson, S. L. 2015. Research methods and statistics: A critical thinking approach. Cengage Learning. Bostan, USA.

Jaya, J., and K. Thanushkodi. 2011. K. Certain investigation on MRI segmentation for the implementation of CAD system. *WSEAS Transaction on Computing* 10 (6):189–98.

Luijnenburg, S. E., D. Robbers-Visser, A. Moelker, H. W. Vliegen, B. J. Mulder, and W. A. Helbing. 2010. Intra-observer and inter-observer variability of biventricular function, volumes and mass in patients with congenital heart disease measured by CMR imaging.

*The International Journal of Cardiovascular Imaging* 26 (1):57–64. doi:10.1007/s10554-009-9501-y.

Marshkole, N., B. K. Singh, and A. S. Thoke. 2011. Texture and shape based classification of brain tumors using linear vector quantization. *International Journal of Computer Applications* 30 (11):21–23.

Menze, B. H., A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, … L. Lanczi. 2015. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging* 34 (10):1993–2024. doi:10.1109/TMI.2014.2377694.

Nery, F., J. S. Silva, N. C. Ferreira, F. Caramelo, and R. Faustino. 2012. An algorithm for the pulmonary border extraction in PET Images. *Procedia Technology* 5:876–84. doi:10.1016/j.protcy.2012.09.097.

Prabha, D. S., and J. S. Kumar. 2016. Performance evaluation of image segmentation using objective methods. *Indian Journal of Science and Technology* 9:1–8.

Prakash, O., M. Khare, and A. Khare. 2013. Image denoising technique based on soft thresholding of biorthogonal wavelet coefficients. *Thresholds* 1 (1):2j.

Saba, L., J. C. Than, N. M. Noor, O. M. Rijal, R. M. Kassim, A. Yunus, R. N. Chue, and J. S. Suri. 2016. Inter-observer variability analysis of automatic lung delineation in normal and disease patients. *Journal of Medical Systems* 40 (6):142. doi:10.1007/s10916-016-0504-7.

Sampat, M. P., Z. Wang, M. K. Markey, G. J. Whitman, T. W. Stephens, and A. C. Bovik. 2006. October. Measuring intra-and inter-observer agreement in identifying and localizing structures in medical images. In *Image Processing, 2006 IEEE International Conference on*, 81–84. Atlanta, GA. doi: 10.1109/ICIP.2006.312367.

Santos, B. S., C. Ferreira, J. S. Silva, A. Silva, and L. Teixeira. 2004. Quantitative evaluation of a pulmonary contour segmentation algorithm in X-ray computed tomography images. *Academic Radiology* 11 (8):868–78. doi:10.1016/j.acra.2004.05.004.

Singh, B. K., K. Verma, L. Panigrahi, and A. S. Thoke. 2017. Integrating radiologist feedback with computer aided diagnostic systems for breast cancer risk prediction in ultrasonic images: An experimental investigation in machine learning paradigm. *Expert Systems with Applications* 90:209–23. doi:10.1016/j.eswa.2017.08.020.

Van Erkel, A. R., and P. M. T. Pattynama. 1998. Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology* 27 (2):88–94. doi:10.1016/S0720-048X(97)00157-5.

World health organization cancer factsheets.[online]:http://www.who.int/mediacentre/factsheets/fs297/en/index.html.