Improved feature extraction and selection method for brain tumor detection

M. Amalmary
Ph.D Research Scholar, PG & Research Department of Computer Science, Hindusthan College of arts and science (Autonomous), Coimbatore, Tamil Nadu, India

Dr. A. Prakash
Professor, PG & Research Department of Computer Science, Hindusthan College of arts and science (Autonomous), Coimbatore, Tamil Nadu, India

Abstract---Medical Image processing is a vital research area in the field of medical and computing technology. Among this Feature extraction and feature selection are the main tasks in pre-processing of info logs to recognize digital protection threats and attacks while using machine learning. With regards to the investigation of heterogeneous data got from various sources, these tasks are figured out to be time-consuming and hard to be overseen effectively. Feature selection and feature extraction techniques have dissected with the end goal of how adequately these techniques can be utilized to accomplish elite of learning algorithms that eventually improves prescient accuracy of classifier. In this phase, Feature Extraction is done by the Wavelet-Based Gray-Level Run-Length Matrix (GLRLM) and the extracted feature set is additionally refined through feature selection. After this Feature Extraction, naive bayes wrapper classifier is used for covering feature selection. Proposed strategy was utilized for feature selection are useful in distinguishing brain tumor where is actually found.

Keywords---Feature Extraction, Wavelet Based Gray Level Run Length Matrix, Pre Processing, Machine Learning.

Introduction

To give better health care to the patient, introduction of Information Technology and e-Health care system in the medical field help the clinical specialists. Unusual development of cells inside the brain occurs, when the Brain tumors influence the humans gravely. It can upset appropriate brain function and be life-threatening. Benign tumors and malignant tumors are the two kinds of brain
tumors. Previous one is less harmful when contrasted and the malignant tumors as malignant are quick creating and harmful while benign are moderate developing and less harmful. The different sorts of medical imaging technologies depend on non-invasive methodology, for example, MRI, CT examines Ultrasound, SPECT, PET and X-beam. Magnetic Resonance Imaging (MRI) is significantly utilized, when distinguished to other medical imaging techniques and it gives more noteworthy difference pictures of the brain and cancerous tissues.

Through MRI images, brain tumor identification can be performed. This work for the most part concentrates on recognizing the brain tumor with the assistance of image processing techniques. The central point of interest here is recognizing the brain tumor at beginning phase and giving the upgraded treatment. Radiological assessment is needed to decide its location, its size, and effect on the encompassing regions, when the brain tumor is clinically suspected. The best therapy, surgery, radiation, or chemotherapy, is decided, in light of this data. In the event that the tumor is perceived precisely in its beginning phase, the odds of endurance of a tumor contaminated patient can be expanded. In this way, in the radiology department, the investigation of brain tumors utilizing imaging modalities has acquired significance. Through image processing, the brain tumor identification is finished.

Improving on the number of resources requested to clarify a colossal arrangement of data precisely is acted in feature extraction. One of the significant issues emerges from the quantity of factors, while playing out the examination of the tedious data. Colossal amount of memory and computation power or an arrangement algorithm requests the training test and sums up inadequately to new examples, while analyzing tremendous factors. Feature extraction is an overall term for techniques for developing mixes of the factors to amend these issues, while clarifying the data with enough precision.

The images that are produced utilizing the MRI images are not clear for diagnosis because of the presence of noise. The noises are added to the image because of different reasons, for example, the conductivity of an item that is exposed to output and imaging (inductive misfortunes). So, after the preprocessing process is done to extricate the subtleties from the crude image that are more pertinent to the tumor forecast process. The feature extraction step is completed for the accompanying reasons: Reduces the class-design variability
Curse of dimensionality
Invariance is accomplished in the ideal structure
To improve the variability in the middleclass design variability.
Features that are extracted are considered to be good only if they have the following characteristics:
(1) Repeatability
(2) Saliency
(3) Efficiency and compactness
(4) A feature should occupy relatively very small area of the image as well as it should be robust to occlusion and clutter.
Feature extraction is to separate the significant information from the signal or image and it helpful to classifier to characterize the image. The feature extraction technique conquers the issue of computational intricacy and over fitting of the prepared data. In this phase Feature Extraction is finished by the Wavelet-Based Gray-Level Run-Length Matrix (GLRLM). The extracted feature set is additionally refined through feature selection. This process picks the subset of features that are extracted based on the standards that suit for the forecast process. Here naive bayes wrapper classifier of covering attribute selection strategy was utilized for feature selection.

**Existing Methodology**

**GLCM Texture Features**

GLCM is texture character profile and this profile delivers to contact for example smooth, plush and unpleasant and so on. The request for character profile statics is: First request texture measures were statistics announced from the original image values, like difference, and pixel neighbor relationship aren't executed. A subsequent request measure depicts the relationship among the gatherings of two (normally neighboring) pixels in the original image. Third and higher request textures (taking note of the relationships between the at least three pixels) are hypothetically conceivable however for all intents and purposes/regularly not executed as a result of computing the time and interpretation difficulty. The connection among the two pixels all at once is gotten by GLCM, which are named as the reference and the neighbor pixel. GLCM elucidates the distance and angular spatial relationship over an image sub-area of a specific size. From gray scale values, we set up the GLCM and it is viewed as how successive the pixel with gray level (gray scale power or gray tone) values come either evenly, vertically and diagonally to leveled the pixels with the value j.

**Bacterial Foraging Optimization (BFO)**

BFO comprises of three head mechanisms in particular, chemo tactic, reproduction, and elimination-dispersal. Chemo tactic (cell development) is the activity of bacteria get-together to nutrient-rich zones in an unconstrained design; in this specific circumstance, a cell-to-cell communication mechanism is established to simulate the biological conduct of bacterial development (swim/tumble). Reproduction comes from the idea of natural selection; under this procedure, simply the best adjusted bacteria will in general endure and send their genetic characters to succeeding generations, while the less adjusted ones will in general perish. Elimination-dispersal occasions haphazardly select pieces of the bacteria populace to diminish and disperse into arbitrary situations in the environment; this way the algorithm ensures the variety of the species, and forestalls getting caught to nearby optima, improving global hunt ability. As in the BFO algorithm, the ideal feature is chosen from each gathering and just those chose features are additionally utilized in the classification.
Proposed Methodology

Feature extraction

The extraction of features from texture district and an edge is known as feature extraction. It is the development of getting the designated graphic mollified from the imaginings for indexing and repossession. While setting aside besides plentiful effort for processing the enormous of datasets hence presume the undesirable features that must be taken out. At that point the select feature contains the pertinent feature of the input image. In this way, feature extraction process based on the exploration work. The fragmented image is the necessary wellspring of the feature extraction. In image classification technique, the major advance is the image features extraction.

In example acknowledgment and in image processing, feature extraction is an uncommon type of dimensionality reduction. At the point when the input data to an algorithm is too huge to possibly be processed and it is suspected to be notoriously excess then the input data will be changed into a decreased portrayal set of (features vector). In this phase Feature Extraction is finished by the Wavelet-Based Gray-Level Run-Length Matrix (GLRLM). This Feature Extraction technique is useful in recognizing brain tumor where is actually found and helps in anticipating next phase. Changing the input data into the arrangement of features is called feature extraction. The extraction of feature is based on the accompanying variables:

- The feature should convey point by point information of the image for extraction and it doesn't need domain-specific knowledge.
- The computation is simple in view of feasibility of assortment of larger images and the fast retrieval.
- The fundamental property of the image is eluded as texture and during the process of recovering the image; it acts a powerful provincial descriptor. It ordered the texture images from the non-textured images and consolidated to another visual characteristic for the process of viable retrieval. The texture extraction made by the structural, statistical, model-based and changes draws near. The feature which depicted the physical composition of a surface.

Feature Extraction Through on the Wavelet-Based Gray-Level Run-Length Matrix (GLRLM):

In this phase, from the start, we apply the wavelet transform to the original image I. Here, we get the four sub bands, which have low-frequency information and high-frequency information like LL, LH, HL, and HH. At that point, we compute the Gray-Level Run-Length Matrix (GLRLM) feature for high-frequency sub bands in light of the fact that the detail coefficients are accessible in the horizontal, vertical, and diagonal sub bands.

Gray-Level Run-Length Matrix (GLRLM)

The Gray Level Run Length technique is a strategy for extricating higher solicitation quantifiable texture data. The quantity of dark levels G, in the image is frequently diminished by re-quantization going before the aggregation of the
matrix (Wang Xinli, 1994). A matrix from of value features may be separated for texture procedure is called as GLRLM. A texture comprehension is a holder of gray power pixel in a specific heading from the reference pixels. The measure of the nearby pixels that have a comparable gray power in a specific heading is named as Run length. Gray level run length matrix is a two dimensional matrix where every constituents \((u, v/ \theta)\) is the number of elements \(v\) by means of the intensity \(u\) in the direction \(\theta\). The Gray Level Run Length matrix is produced as,

\[
K(\theta) = (r(u, v)/ \theta), \ 0 \leq u \leq N_r, 0 \leq v \leq Kmax \quad (1)
\]

Where \(N_r\) the superlative is gray level, \(k\) max is maximal length along with \(u, v\) is a matrix size values.

GLRLM accomplishes more data by using the gray force level worth co-props value. These statistics gives information about the texture part of an image. These are as follows,

- Short Run Emphasis (SRE)
- Long Run Emphasis (LRE)
- Gray-Level Non uniformity (GLN)
- Run Length Non uniformity (RLN)
- Run Percentage (RP)

(a) Short Run Emphasis (SRE)

\[
SRE = \frac{1}{n_g} \sum_{u=1}^{R} \sum_{v=1}^{S} \frac{z(u,v)}{v^2} \quad (2)
\]

The SRE is greatly reliant on the appearance of short runs and is anticipated large for excellent textures.

(b) Long Run Emphasis (LRE)

Long Run Emphasis measures distribution of long runs. The LRE is very reliant on the presence of long runs and is suspected large for abrupt structural textures.

\[
LRE = \frac{1}{n_g} \sum_{u=1}^{R} \sum_{v=1}^{S} z(u,v) \ast v^2 \quad (3)
\]

(c) Gray-Level Non uniformity (GLN)

GLN measures the resemblance of gray level guidelines all through the image. The GLN is predictable little if the gray level principles are similar all through the image.

\[
GLN = \frac{1}{n_g} \sum_{u=1}^{R} (\sum_{v=1}^{S} z(u,v))^2 \quad (4)
\]

(d) Run Length Non uniformity (RLN)

RLN measures the length’s similitude of continues to run all through the image; The RLN is predictable little if the run lengths are indistinguishable all through the image.

\[
RLN = \frac{1}{n_g} \sum_{u=1}^{S} \sum_{v=1}^{R} (z(u,v))^2 \quad (5)
\]
(e) Run Percentage (RP)

The homogeneity just as the circulation of convey running of a picture in an exact direction the RP is the frequent when the length of runs is one for all gray levels specifically path.

\[ RP = \frac{n_g}{z(u,v)*v} \quad (6) \]

Maximum Intensity (MI)

The histogram of an image is generally identified with a pixel’s histogram compel values regarding the image processing association. Moreover, it tends to a diagram depicting the quantities of pixels in an image at every single extraordinary oblige worth found in the relative image. Since there are 256 diverse realizable powers for a 8-cycle grayscale image, the histogram graphically recommends 256 numbers depicting the assignment of pixels among those grayscale values.

\[ I = \max_{i=0}^{255}(\text{count of pixelintensity}) \quad (7) \]

A run-length is represented by \((gl, rl, θ)\), where \(gl\), \(rl\), and \(θ\) are the gray level, run-length, and direction respectively. The run-length matrix is a method of looking through the image, consistently across a provided guidance, for runs of pixels having a similar gray level worth. Accordingly, provided guidance, the run-length matrix measures each permitted gray level an incentive into how often there are runs of, for instance, 2 consecutive pixels with a similar worth. Next it does likewise for 3 consecutive pixels, at that point for 4, 5, etc. Gray level run-length features are a type of gray level statistical features. Note that a wide range of run-length matrices might be registered for a single image, one for each picked direction.

\[ R(i,j) = (g(i,j)|i), 0 \leq i \leq N_g, 0 \leq j \leq R_{max} \quad (8) \]

Where \(N_g\) is the maximum gray level and \(R_{max}\) is the maximum length. Six texture features can be removed from the GLRLM. These features utilize gray level of pixel in arrangement and is planned to distinguish the texture that has a similar estimation of SRE and LRE however have differences in the distribution of gray levels.

Algorithm for Feature Extraction

| Input: Img: Image  
GLRLM: Gray Level Run Length Matrix  
Output: X[N : M]: Feature Matrix  
Step 1: Set θ as direction parameter, S and M denotes the feature descriptor & number of features respectively.  
Step 2: Initialize P as number of directions (4: 0°, 45°, 90° and 135°)  
Step 3: Assign S  
Step 4: M ← P × S |
Step 5: for $i = 1$ to Image do
Step 6:  Determine the GLRL matrix using grayrlm( ) for the input image (ip)
Step 7:  for $q = 1$ to $p$ do
Step 8:  $GLRM_{\theta_q} = grayrlm (ip, \theta_q )$
Step 9:  for $x = 1$ to $S$ do
Step 10: compute the $GLRM_{\theta_q}$ and append it to $X$
Step 11: end for
Step 12: end for
Step 13: end for
Step 14: return $X$

Feature Selection

The extracted feature set is additionally refined through feature selection. This process picks the subset of features that are extracted dependent on the standards that suit for the prediction process. To decrease the intricacy of classification feature selection ought to perform. In these above automated chosen features are distinguished utilizing naive bayes classifier of wrapper quality selection strategy.

A. Naive bayes wrapper classifier

Naive bayes classifier is one of the probabilistic classifier in wrapper quality selection. The picture feature data is numeric data. For numeric data, the suitable classifier is either bayes network, choice trees. The classification depends on probability of features accessible. Feature selection is likewise done utilizing naïve bayes classifier. They chose features are appeared in Figure 1.

![Figure 1. Selected improved features using naive bayes classifier](image)

Among all the extricated features, chosen texture features are upgraded features as shading, texture and edge. The recognized features are addressing the picture characteristics for classification.
Bayesian Network (BN)

Bayesian organization (BN) (Pearl, 1988) comprises of a coordinated acyclic graph $G$ and a set $P$ of probability distributions, where nodes and bends in $G$ address random variables and direct correlations between variables individually, and $P$ is the arrangement of local distributions for every hub. A local appropriation is ordinarily indicated by a conditional probability table (CPT). BNs are of regularly utilized for the classification issue. In the classification learning issue, a student endeavors to develop a classifier from a given arrangement of marked preparing models that are addressed by a tuple of trait variables utilized on the whole to predict the estimation of the class variable. The significant undertakings are to gather the construction of the organization from the given picture data and make the conditional probability table (CPT). The Pseudo code for classification of picture quality data by utilizing Bayes net is:

**Step 1:** Select Estimator algorithm as basic Estimator having the min probability of 0.5 for finding the contingent probability tables of the Bayes Network.

**Step 2:** Select an algorithm utilized for looking through network structures as slope climbing algorithm. With suspicions as max no. of parents as 2, random request as false, use AD tree as off.

Every one of the nodes a probability distribution for the node given its parents is specified too. For instance, in the Bayes net underneath there is a restrictive distribution for homogeneity given the estimation of homogeneity. Since homogeneity has no parents, there is an unequivocal distribution for homogeneity. The performance of the classes is defined by utilizing true positive, false positive, true negative and false negative qualities. The performance of the classes is defined by utilizing true positive, false positive, true negative and false negative qualities.

After the performance of classes are assessed by utilizing three parameters precision, recall and F-measure.

$$\text{Precision} = \frac{tp}{tp - fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

The F-measure, initially presented in information recovery, is these days regularly utilized as a performance metric for issues like binary classification, multi-mark classification, and structured output prediction. F-measure is the harmonic mean of precision and recall.

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
Experimental Result

Precision

Table 1. Comparison Table of Precision Ratio

<table>
<thead>
<tr>
<th>No of Images</th>
<th>GLCM</th>
<th>BPO</th>
<th>IFE&amp;FSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.42</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>150</td>
<td>0.49</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td>225</td>
<td>0.52</td>
<td>0.43</td>
<td>0.57</td>
</tr>
<tr>
<td>300</td>
<td>0.59</td>
<td>0.48</td>
<td>0.61</td>
</tr>
<tr>
<td>375</td>
<td>0.55</td>
<td>0.51</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The Comparison table 1 of Precision of Value explains the different values of existing algorithm GLCM, BPO and proposed improved IFE&FSM Algorithm. While comparing the Existing algorithm, the proposed improved IFE&FSM provides the better results. The existing algorithm (GLCM, BPO) precision values start from 0.42 to 0.55, 0.37 to 0.51 and proposed IFE&FSM Algorithm starts from 0.48 to 0.65, provides the great results.
The comparison chart of Precision values demonstrates the existing GLCM, BPO and proposed IFE&FSM algorithm. The proposed algorithm is better than the existing algorithm. X axis denotes the number of images and Y axis denotes the Precision Ratio. The existing algorithms (GLCM, BPO) precision values start from 0.42 to 0.55, 0.37 to 0.51 and proposed IFE&FSM Algorithm starts from 0.48 to 0.65, provides the great results.

**Recall Ratio**

<table>
<thead>
<tr>
<th>No of Images</th>
<th>GLCM</th>
<th>BPO</th>
<th>IFE&amp;FSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.44</td>
<td>0.38</td>
<td>0.46</td>
</tr>
<tr>
<td>150</td>
<td>0.53</td>
<td>0.46</td>
<td>0.54</td>
</tr>
<tr>
<td>225</td>
<td>0.58</td>
<td>0.52</td>
<td>0.59</td>
</tr>
<tr>
<td>300</td>
<td>0.62</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>375</td>
<td>0.64</td>
<td>0.62</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The Comparison table 2 of Recall of Value explains the different values of existing algorithm GLCM, BPO and proposed improved IFE&FSM Algorithm. While comparing the Existing algorithm, the proposed improved IFE&FSM provides the better results. The existing algorithm (GLCM, BPO) values start from 0.44 to 0.64, 0.38 to 0.62 and proposed IFE&FSM Algorithm starts from 0.46 to 0.65, provides the great results.

The comparison chart of Recall values demonstrates the existing GLCM, BPO and proposed IFE&FSM algorithm. The proposed algorithm is better than the existing algorithm. X axis denotes the number of images and Y axis denotes the Recall Ratio. The existing algorithms (GLCM, BPO) values start from 0.44 to 0.62 and proposed IFE&FSM Algorithm starts from 0.46 to 0.65, provides the great results.
F- Measure

Table 3. Comparison Table of F- Measure

<table>
<thead>
<tr>
<th>No of Images</th>
<th>GLCM</th>
<th>BPO</th>
<th>IFE&amp;FSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.49</td>
<td>0.39</td>
<td>0.56</td>
</tr>
<tr>
<td>150</td>
<td>0.54</td>
<td>0.48</td>
<td>0.59</td>
</tr>
<tr>
<td>225</td>
<td>0.56</td>
<td>0.52</td>
<td>0.62</td>
</tr>
<tr>
<td>300</td>
<td>0.63</td>
<td>0.59</td>
<td>0.67</td>
</tr>
<tr>
<td>375</td>
<td>0.67</td>
<td>0.64</td>
<td>0.69</td>
</tr>
</tbody>
</table>

The Comparison table 3 of F-Measure of Value explains the different values of existing algorithm GLCM, BPO and proposed improved IFE&FSM Algorithm. While comparing the Existing algorithm, the proposed improved IFE&FSM provides the better results. The existing algorithm (GLCM, BPO) values start from 0.49 to 0.64 and proposed IFE&FSM Algorithm starts from 0.56 to 0.69, provides the great results.

![Figure 5. Comparison Chart of F-Measure](image)

The comparison chart of Recall values demonstrates the existing GLCM, BPO and proposed IFE&FSM algorithm. The proposed algorithm is better than the existing algorithm. X axis denotes the number of images and Y axis denotes the F-Measure Ratios. The existing algorithms (GLCM, BPO) values start from 0.49 to 0.64 and proposed IFE&FSM Algorithm starts from 0.56 to 0.69, provides the great results.

Weighted Average of all attributes

Table 4. Weighted Average of all attributes

<table>
<thead>
<tr>
<th></th>
<th>COUNT</th>
<th>HOMOGENEITY</th>
<th>ENERGY</th>
<th>CORRELATION</th>
<th>CONTRAST</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>1.2</td>
<td>0.5</td>
<td>1.3</td>
<td>2.2</td>
<td>2.6</td>
<td>3.4</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Precision</td>
<td>1.3</td>
<td>0.7</td>
<td>1.5</td>
<td>2.3</td>
<td>2.7</td>
<td>3.5</td>
</tr>
<tr>
<td>Recall</td>
<td>1.5</td>
<td>1.9</td>
<td>2.2</td>
<td>2.6</td>
<td>3.1</td>
<td>3.7</td>
</tr>
</tbody>
</table>
The comparison chart of Figure 6 values demonstrates the performance of Weighted Average of all attributes (ROC Area, F-Measure, Recall, Precision, FP Rate, and TP Rate). Among all the features (Count, Homogeneity, Energy, Correlation, and Mean), Homogeneity is showing the high performance with the better result of ROC Area.

**Conclusion**

In this phase proposed improved feature extraction and feature selection method. In this phase Feature Extraction is finished by the Wavelet-Based Gray-Level Run-Length Matrix (GLRLM). The extracted feature set is additionally refined through feature selection. This process picks the subset of features that are extracted based on the standards that suit for the forecast process. Here naive bayes wrapper classifier of covering attribute selection strategy was utilized for feature selection. Experimental result proved that the proposed provides better performance than existing methods. This Improved Feature Extraction and Feature Selection techniques are useful in distinguishing brain tumor where is actually found and helps in anticipating next phase.

**References**


13421


