BONAFIDE CERTIFICATE

Certified that this project report, "**Brain Cancer Detection and Staging Using Machine Learning**" is the bonafide work done by **MUHAMED Z [19EC1063], KATTA SATHVIK [19EC1047] and VIJAYA KRISHNA TEJA [19EC1022]** Eight semester B.Tech. class of **Electronics and Communication Engineering** in the project work (phase I) during the year 2022-2023.

PROJECT GUIDE	**HEAD OF THE DEPARTMENT**
Dr. S. BATMAVADY	**DR. V. SAMINADAN**

Submitted for the practical exam held on ________________________________

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

The successful completion of the project is indeed practically incomplete without the mention of all those people who greatly supported and encouraged us throughout the project. We feel grateful to our guide **Dr. S. BATMAVADY**, Professor, Department of Electronics and Communication Engineering, Puducherry Technological University, for her encouragement and support. She has been a source of valuable guidance, suggestions and kindness during the project work.

We would like to express our sincere thanks to our **VICE CHANCELLOR Dr. S. MOHAN,** Puducherry Technological University for providing the college facilities for the completion of this project.

We would like to express our heartfelt gratitude to **Dr. V. SAMINADAN,** Professor and Head of the Department of Electronics and Communication Engineering, Puducherry Technological University for providing all department facilities and guidance which enabled us to complete the project.

We feel obliged to thank the review by the panel members **Dr. L. NITHYANANDAN** Associate Professor, **Dr. D. SARASWADY** Professor, **Dr. K. KUMAR** Professor, of the Electronics and Communication Engineering Department, for their support, encouragement and guidance.We express our deep sense of gratitude to all teaching and non-teaching staff of our department, and to our friends for their support and encouragement during the entire course of this dissertation work. We also express our sincere thanks to our parents who motivated us in all means.

ABSTRACT

Now a days , stage identification of Brain cancer became a challenging issue for pulmonologist form the Computed Tomography (CT) scan directly. An early identification in Brain cancer and corresponding medication will save a person from sick. Machine learning algorithms along with image processing techniques help doctors for early identification of cancer .Cancer cells from different organs of the body are carried in blood to the Brain where they get stagnated and develop into a tumor .In this work, we proposed a precise algorithm to identify tumor regions from a CT scan of Brain .Since a CT scan is affected with Gaussian and salt – pepper noise a median filter is best suited for removing the noise in the pre-processing step .We converted the grayscale image into binary image for further processing .Watershed algorithm is a new kind of segmentation process where the segmentation is carried out by identifying the catchment basins and segregate the regions accordingly. Tumors are extracted from morphological operations viz., opening and erosion .Under feature extraction, features like area, perimeter, eccentricity and diameter are extracted for the stage identification .We used a multi class Support Vector Machine (SVM) classifier with a data set containing records of 2574 patients which gave us the best possible classification according to stage with less time - consumption .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2-D	Two-Dimensional
AHE	Adaptive Histogram Equalization
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CCPA	California Consumer Privacy Act
CDF	Cumulative Distribution Function
CIA	Cancer Imaging Archive
CLAHE	Contrast-limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CT	Computed Tomography
EISPACK	Eigen System Package
FAQs	Frequently Asked Questions
GDPR	General Data Protection Regulation
GM	General Motors
LINPACK	Linear System Package
MATLAB	Matrix Laboratory
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PCA	Principal Component Analysis
PII	Personally Identifiable Information
RBF	Radial Basis Function
RGB	Red, Green and Blue
SVD	Singular Value Decomposition
SVM	Support Vector Machine
X-Ray	X-radiation

CHAPTER 1

INTRODUCTION

1.1 PRELUDE

Cancer is one of the deadliest diseases around the globe. The major kinds of cancers are carcinoma – cancer that forms in epithelial tissue, sarcoma – cancer that forms in bones, melanoma – cancer that forms in skin, lymphoma – cancer that forms in lymphatic system, leukemia – cancer that forms in blood tissues, etc. The cause of brain cancer is still largely unknown. Although some genetic conditions and environmental factors may contribute to the development of brain cancer, the risk factors are much less defined for brain cancer than for other cancers in the body. Blood reaches to Brain from various parts of the body like brain, breast, etc., where the osmosis process takes place for purification. Brain cancer can be diagnosed with Magnetic Resonance Imaging(MRI) and CT scans. As there are some advantages of CT scan over MRI like cost economic and lower level of risk, It is best advised that CT scans can be used for Brain cancer diagnosis. Medical image processing was became an evolutionary break through form 1960's which was developed and upgraded at every stage and helps in small textures of tumors and tissues too .So, Image processing techniques can be a great idea for choosing to detect Brain cancer .Fig.1 shows the CT scan image of Brain .A CT scan image seems to be a grayscale image but it is actually not. It composed of red, green, blue planes. It should be converted to a grayscale image since processing of color image takes huge memory and high processing time. The CT image is mainly affected with salt and pepper noise which can be easily removed by median filter. The filtered output should be converted to a binary image to extract tumor regions. Watershed segmentation is the core part in this project where the image is segmented on basis of catchment basin. An image is considered to be topological surface and the shallow regions are grouped and segmented accordingly. We identified and rectified the drawbacks of watershed segmentation. Morphological operations including masking operations is are used to extract exact tumor regions . Image processing techniques have to face some difficulties in deciding the structure that was detected. It can be a tumor or blood tissue. Features like area, perimeter, centroid and eccentricity are extracted from region-based feature extraction for the tumors that were extracted from segmentation. The stage of the Brain cancer can be determined by the size of the tumor that grown. Based on the features extracted from the tumor structures, the radius of the tumor mor an easily be deducted and stage of the tumor can be classified according to the radius of the tumor. Support Vector machine (SVM) is a supervised machine learning algorithm that is used to classifying a new input based on the pretrained model. The hyperplane in the SVM is the plane passing midway through the support vectors.

1.2 MOTIVATION

With today's technology, doctors can replace every part of the human body, from bones to organs, hands, and face except lung and Brain. Hence early detection of damage in the Brain or lung should be recognized to improve the survival rate of human beings. There are many techniques to diagnose Brain cancer such as Chest Radiograph (X-Ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), etc. But even after analyzing these reports, doctors may not accurately predict the cancer stage or tumor size. Therefore, there is a great need for new technology, i.e. Image Processing Techniques, which is an excellent tool to improve manual analysis and to predict the size of tumor cells more accurately.

1.3 LITERATURE REVIEW

Table 1.1 Literature Review

S.NO	AUTHOR	TITLE OF THE JOURNAL	PROPOSED TECHNIQUE	LIMITATIONS
1	Imran Shafi, Sadia Din	An Effective Method for Brain Cancer Diagnosis from CT Scan Using Deep Learning	CNN, SVM	Accuracy is 92%
2	Nikitha Benarjee, Subbalaxmi Das	Prediction Brain Cancer – In Machine Learning Perspective	SVM, ANN, RANDOM FOREST	Accuracy is 90%
3	Azmira Krishna, P.C Srinivasa Rao	Computerized Classification of CT Brain Images using CNN	CNN, SVM, BPNN	Accuracy is 86.5%
4	Arun B. Mathews , Dr.M.K. Jeyakumar	Analysis of Brain Tumor Detection using Various Segmentation Techniques	k-Nearest Neigbour, SVM, Decision Tree, MLP	Accuracy is 88.55%
5	Sanjukta Rani Jena, Dr. Thomas George	Feature Extraction And Classification Techniques For The Detection of Brain Cancer: A Detailed Survey	MLP, Neural Network, Gradient Tree	Accuracy is 90%
6	Alakawa	Brain Cancer Detection for Multi-Class SVM Approach	Multi-class SVM classifier	97% for cancer identification and 87% for cancer prediction

1.4 OBJECTIVE

To enhance the identification of Brain cancer tissues with better accuracy and to classify the stage of the cancer.

1.5 METHODOLOGY

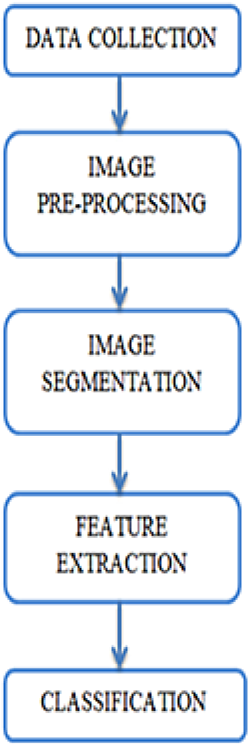


Figure 1.1 Methodology

First the CT image undergoes pre-processing. After preprocessing the processed image is segmented using watershed segmentation. This gives the image with cancer nodules marked. In addition to features like area, perimeter and eccentricity, features like Centroid, Diameter and Pixel Mean Intensity have been extracted in feature extraction stage for the detected cancer nodules. Therefore, additional stage of classification of cancer nodule has been performed using Multiclass Support Vector Machine. Extracted features are used as training features and trained model is generated. Then, unknown detected cancer nodule is classified using that trained prediction model.

- Image Preprocessing

  Firstly, in image pre-processing median filter is used on grayscale image of CT scan images. Some noises are embedded on CT Images at the time of image acquisition process which aids in false detection of nodules. Noise may be detected as cancer nodules sometimes. Therefore, these noises have to be removed for accurate detection of cancer. Median filter removes salt and pepper noise from the CT images. After median filter, Gaussian filter is implemented. It smoothens the image and removes speckle noise from image.

- Segmentation

  This process locates objects or boundaries which help in acquiring the region of interest in the image. It partitions the image into regions to identify the meaningful information. In Brain cancer detection it segments the cancer nodule from the CT scan image. In the proposed model watershed segmentation is implemented. Its main feature is that it can separate and identify the touching objects in the image. This feature helps in proper segmentation of cancer nodules if it is touching to other false nodules.

- Features extraction

  In this stage, features like area, perimeter, centroid, diameter, eccentricity and mean intensity. These features later on are used as training features to develop classifier.

- Classification

  This stage classifies the detected nodule as stage I, II, III, IV. Multiclass Support vector machine (SVM) is used as classifier. It is supervised machine learning method. SVM defines the function that classifies data into four classes.

1.6 ORGANISATION OF THESIS

Chapter 1 introduces the project and elaborates on the literature review, motivation, and project objective

Chapter 2 provides a project description, block diagram, project flow.

Chapter 3 contains the software and technique used

Chapter 4 contains about the SVM Algorithm

Chapter 5 project result, conclusion and future scope

CHAPTER 2

PROJECT DESCRIPTION

2.1 EXISTING SYSTEM

In Existing technology SVM classifies the data points into two classes of hypothesis function i.e. cancerous and non-cancerous. CNN is used for feature selection and the machine learning model SVM is used for classification. CNN selects the important features by max-pooling layer and classifies those features using a fully connected layer. However, the max-pooling layer is inefficient in preserving spatial information and, therefore, loses the information and the stages of Brain cancer were not inferred.

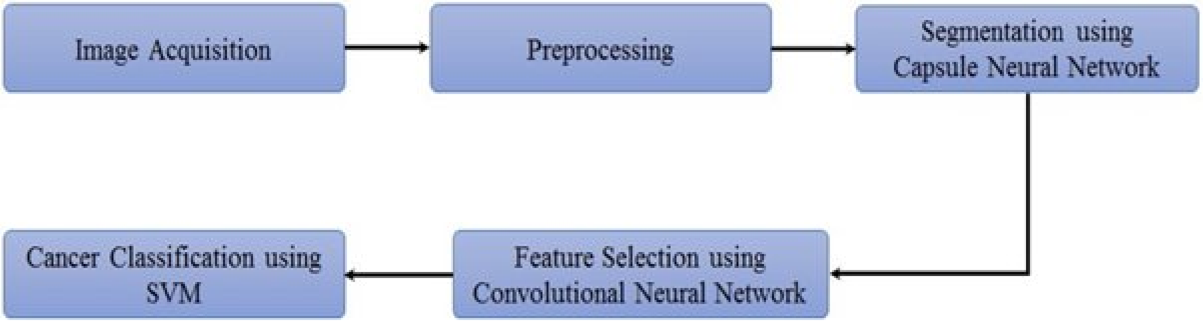


Image Acquisition
Preprocessing
Segmentation using Capsule Neural Network
Feature Selection using Convolutional Neural Network
Cancer Classification using SVM

Figure 2.1 Existing System Methodology

2.2 PROPOSED SYSTEM

Changes on current best solution have been made and new model has been proposed. After preprocessing the processed image is segmented using watershed segmentation. This gives the image with cancer nodules marked. In addition to features like area, perimeter and eccentricity, features like Centroid, Diameter and Pixel Mean Intensity have been extracted in feature extraction stage for the detected cancer nodules. Therefore, additional stage of classification of cancer nodule has been performed using Multiclass Support Vector Machine. Extracted features are used as training features and trained model is generated. Then, unknown detected cancer nodule is classified using that trained prediction model and performance metrics are calculated.

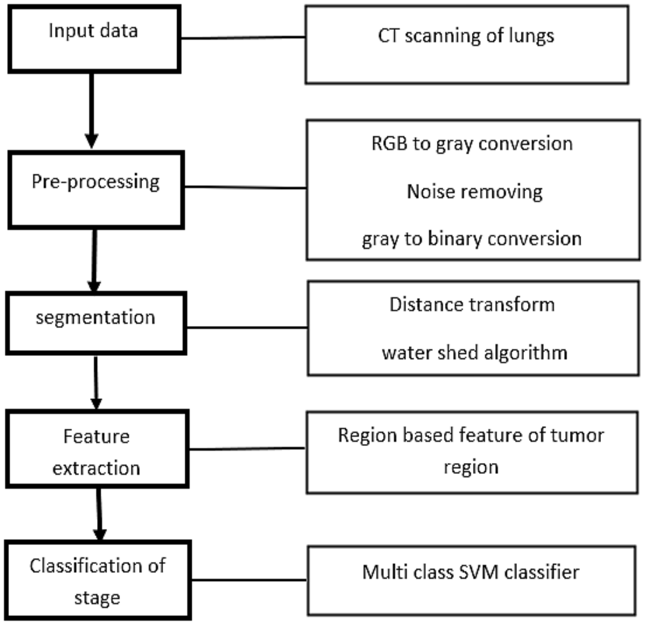


Figure 2.2 Proposed System Methodology

2.2.1 Data Collection

The images which are downloaded from the Internet are shown below. In the following steps, these Brain cancerous images are used for performing different operations to get the desired result(stage of cancer).

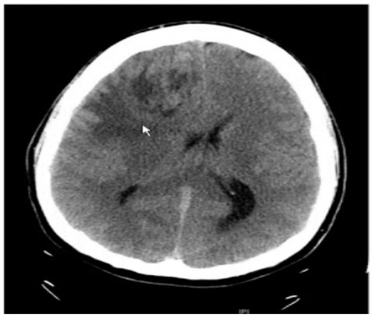


Figure 2.3 Image 1

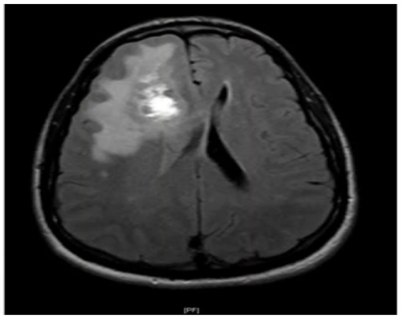


Figure 2.4 Image 2

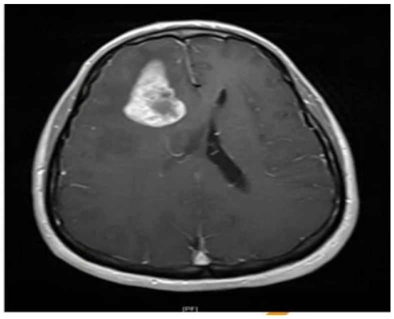


Figure 2.5 Image 3

2.2.2 Pre-Processing

Image pre-processing is the term for operations on images at the lowest level of abstraction. These operations do not increase image information content but they decrease it if entropy is an information measure. The aim of pre-processing is an improvement of the image data that suppresses undesired distortions or enhances some image features relevant for further processing and analysis task.

Image preprocessing consists of three steps to enhance the image quality for further processing.

2.2.2.1 RGB to GRAYSCALE Conversion

Pre-processing step includes RGB to gray conversion, Noise removal and conversion of grayscale to binary image. A RGB image is converted to a grayscale image by luminosity method where the gray scale pixel value is calculated from the following formula;

$$G= ((0.3*R) + (0.59*G) + (0.11*B))$$

Where G = output gray level value

R = Intensity of RED plane

G = Intensity of GREEN plane

B = Intensity of BLUE plane

This implies that red includes 30%, Green includes 59% and blue includes 11% based on the wavelengths. Since RED has the highest wave length it includes most of the black region in the conversion so, its effect should be reduced.

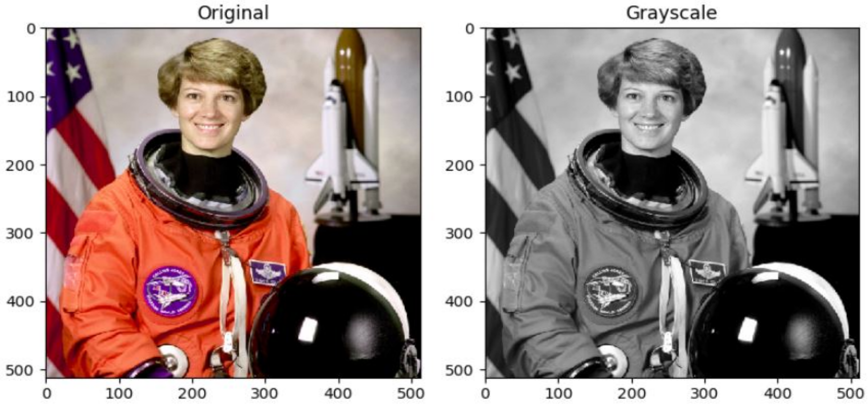


Figure 2.6 Grayscale Converted Image

2.2.2.2 Median Filter

The median filter is the filtering technique used for noise removal from images and signals. Median filter is very crucial in the image processing field as it is well known for the preservation of edges during noise removal. The prior duty of the filter is to scan every input data interceding the overall entries with the median function known as “window” method. The window tends to be a little bit complex over the higher-dimensional signals. The number of medians is set according to the number of windows, falling under odd and even categories.

Median filter is one of the well-known order-statistic filters due to its good performance for some specific noise types such as “Gaussian,” “random,” and “salt and pepper” noises. According to the median filter, the center pixel of a $M \times M$ neighborhood is replaced by the median value of the corresponding window. Note that noise pixels are considered to be very different from the median. Using this idea median filter can remove this type of noise problems . We use this filter to remove the noise pixels on the protein crystal images before binarization operation.

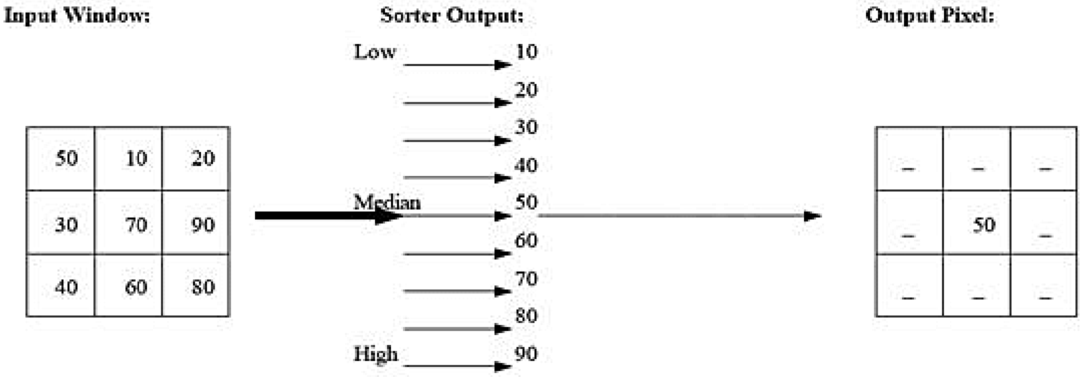


Figure 2.7 Median Filter

Consider a 3x3 image as shown;

Table 2.1 Image Values

0	10	33
65	255	100
155	1	173

Where the pixels 0 and 255 indicates the salt and pepper noise equivalents

Extending border values outside with values at boundary

Table 2.2 Extended Values

0	0	10	33	33
0	0	10	33	33
65	65	255	100	100
155	155	1	173	173
155	155	1	173	173

Median operation on a pixel results in the middle value in the sorted list of neighborhoods of that pixel.

1. Neighborhoods of the pixel 0 in the 8 connectivity are,

    {0,0,10,10,0,0,10,65,65,255}

    The median of the above set of values is 10. So, 0 is replaced with 10.

2. Neighborhoods of the pixel 255 in the 8 connectivity are,

    {0,10,33,65,255,100,155,1,173}

The median of the above set of values is 65. So, 255 is replaced with 65.

Similarly, On applying the above mathematical algorithm on all the pixels, the resultant image is.,

Table 2.3 Resultant Values

10	33	33
65	65	100
155	155	173

It is seen that the salt and pepper is removed and the quality of the image is enhanced.

2.2.2.3 Histogram Equalization

A histogram of an image is the graphical interpretation of the image's pixel intensity values. It can be interpreted as the data structure that stores the frequencies of all the pixel intensity levels in the image.

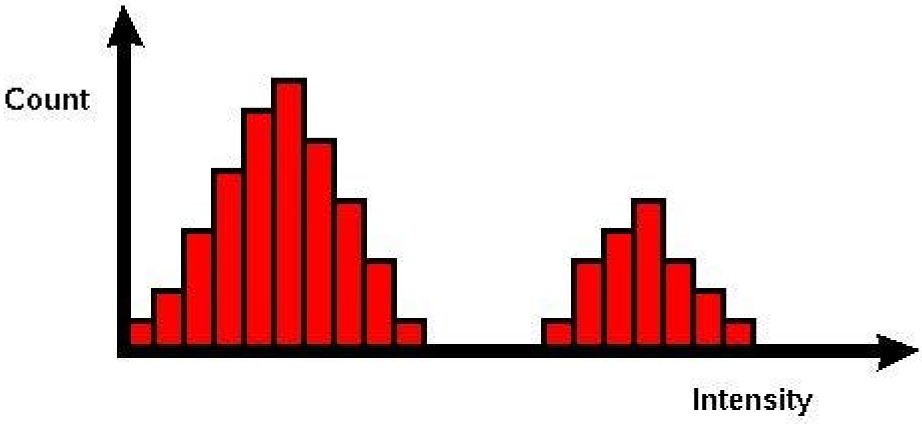


Figure 2.8 Histogram

From the above figure, X-axis represents the pixel intensity levels of the image. The intensity level usually ranges from 0 to 255. For a gray-scale image, there is only one histogram, whereas an RGB colored image will have three 2-D histograms — one for each color. The Y-axis of the histogram indicates the frequency or the number of pixels that have specific intensity values.

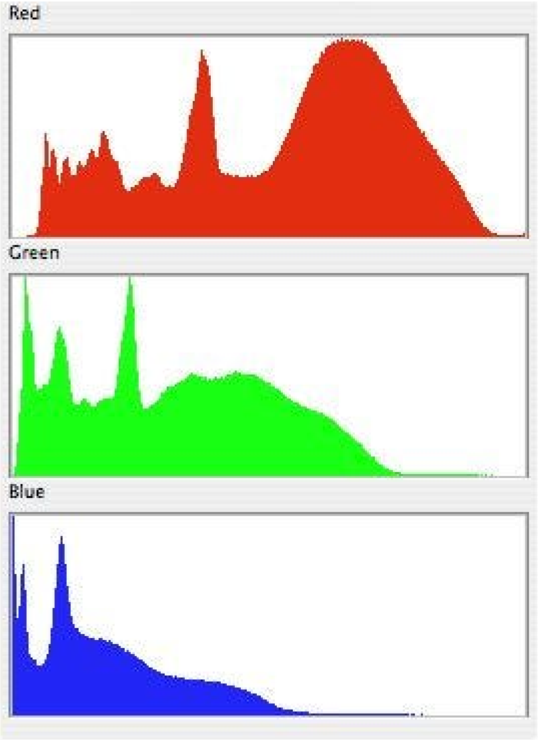


Figure 2.9 RGB

Histogram Equalization is an image processing technique that adjusts the contrast of an image by using its histogram. To enhance the image's contrast, it spreads out the most frequent pixel intensity values or stretches out the intensity range of the image. By accomplishing this, histogram equalization allows the image's areas with lower contrast to gain a higher contrast.

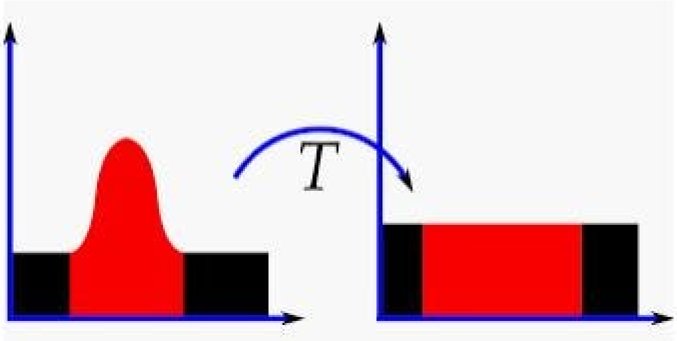


Figure 2.10 Histogram Equalization

In addition to the ordinary histogram equalization, there are two advanced histogram equalization techniques called -

1. Adaptive Histogram Equalization
2. Contrastive Limited Adaptive Equalization

Unlike ordinary histogram equalization, adaptive histogram equalization utilizes the adaptive method to compute several histograms, each corresponding to a distinct section of the image. Using these histograms, this technique spread the pixel intensity values of the image to improve the contrast. Thus, adaptive histogram equalization is better than the ordinary histogram equalization if you want to improve the local contrast and enhance the edges in specific regions of the image.

Contrastive limited adaptive equalization (CLAHE) can be used instead of adaptive histogram equalization (AHE) to overcome its contrast overamplification problem. In CLAHE, the contrast implication is limited by clipping the histogram at a predefined value before computing the CDF. This clip limit depends on the normalization of the histogram or the size of the neighborhood region. The value between 3 and 4 is commonly used as the clip limit.

2.2.2.4 GRAY to BINARY Conversion

We further process the image by converting it into a binary image because, it takes less time for processing and storage. In this work we kept a global threshold value for conversion from the histogram of the image.

$$K = M + 0.5 * SD$$

Where K = Threshold Value

SD = Standard Deviation

M = Mean

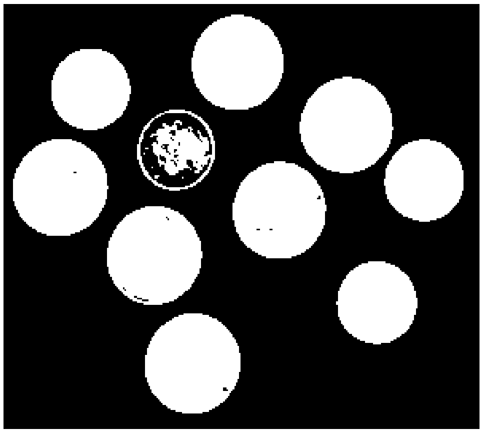


Figure 2.11 Binary Image

2.2.3 Image Segmentation

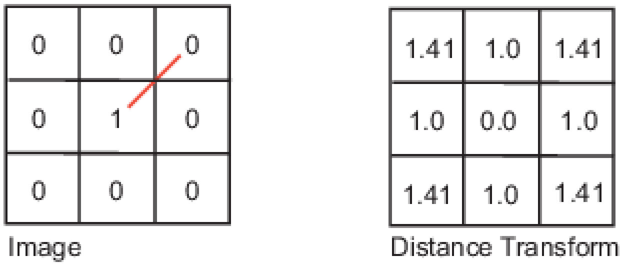
Watershed segmentation is a gradient-based segmentation technique. It considers the gradient map of the image as a relief map. It segments the image as a dam. The segmented regions are called catchment basins. Watershed segmentation solves a variety of image segmentation problem. It is suitable for the images that have higher intensity value. Watershed segmentation is caused over segmentation. To control over segmentation, marker controlled watershed segmentation is used. Sobel operator is suitable for edge detection. In marker controlled watershed segmentation, sobel operator is used to distinct the edge of the object.

We proceeded further to segmentation where we used distance transform-based watershed segmentation

D = bwdist(BW) computes the Euclidean distance transform of the binary image BW. For each pixel in BW, the distance transform assigns a number that is the distance between that pixel and the nearest nonzero pixel of BW. bwdist uses the Euclidean distance metric by default. BW can have any dimension. D is the same size as BW.

[D,L] = bwdist(BW) also computes the nearest-neighbor transform and returns it as label matrix L, which has the same size as BW and D. Each element of L contains the linear index of the nearest nonzero pixel of BW.

[D,L] = bwdist(BW,METHOD) computes the distance transform, where METHOD specifies an alternate distance metric. METHOD can take any of these values:



0	0	0
0	1	0
0	0	0

Image

1.41	1.0	1.41
1.0	0.0	1.0
1.41	1.0	1.41

Distance Transform

Figure 2.12 Distance Transform Values

Generally, an image is considered to be a topological surface assuming the pixel intensities are the heights at different regions. Watershed algorithm finds the two regions namely, Catchment basin and Rigid lines. Distance transform applied on a binary image result in a gray scale image. It calculates the nearest white pixel for each pixel. In binary image, there are only two pixels values viz., 0 and 1. Where 0 resembles black and 1 resembles white. So, all the pixels having intensity 1 will be replaced with 0 and all the pixels with intensity 0 will be replaced with the distance measured from that pixel to nearest pixel with intensity 1. While measuring the distance, we considered 8-connectivity between the pixels. So, segmentation is performed along the Rigid lines of the complemented distance transformed image.

The main disadvantage of watershed segmentation is over-segmentation, which occurs when the surface area is over-segmented into many small, shallow, and insignificant hills and dales features rather than a few large hills/dales features. H-minima transform on the distance-transformed image can prevent it.

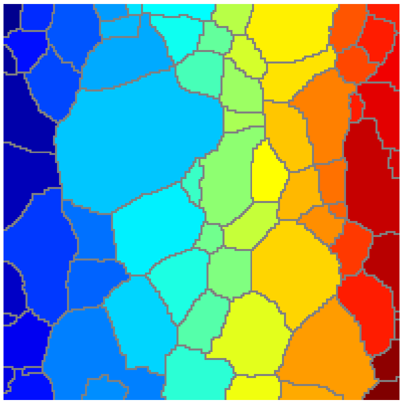


Figure 2.13 Image Segmentation

We can observe that, the tumor regions are more accurately segmented with H-minima transform than the normal segmentation.

2.2.4 Feature Extraction

Boundary tracing followed by labelling the tumor regions is performed on the segmented output for better interpretation of the tumor regions. In Feature extraction, some of the important region-based features or parameters of the image like area in mm^2, perimeter in mm, centroid, eccentricity and orientation are extracted using region props. From the radius obtained, the stage of the cancer is estimated by the radiologists.

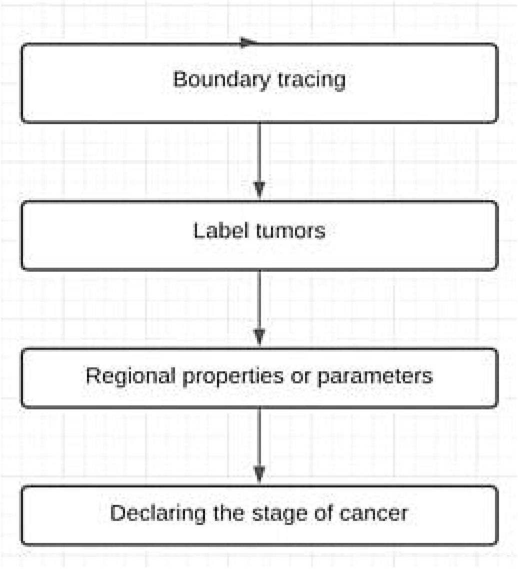

Boundary tracing
Label tumors
Regional properties or parameters
Declaring the stage of cancer

Figure 2.14 Feature Extraction

2.2.5 Classification

Further in processing, used a machine learning algorithm to classify the tumor regions according to their stages. Support vector Machine (SVM) with Watershed segmentation gives the best possible accuracy. SVM is supervised machine learning algorithm and is used for binary classification in general. Adding to that, designed a SVM model to classify in 4 different classes in accordance with 4 different stages of cancer. Fig 10. shows the network architecture of the SVM algorithm

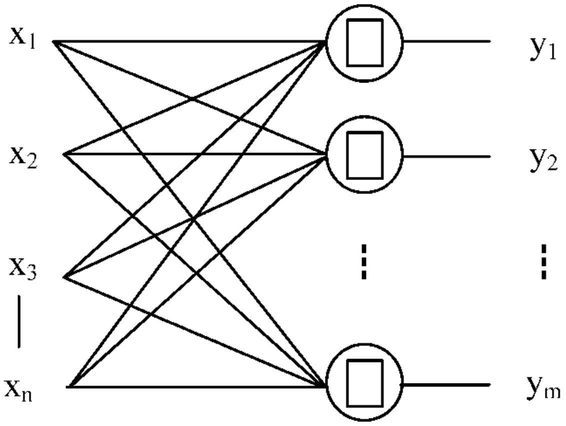


Figure 2.15 Internal networks of Multi Class SVM

Here, used four(4) different classes resembling the four stages of Brain cancer according to the radius of the tumor obtained from the feature extraction.

Table 2.4 Stages

Radius (r)	Stage
r<3mm	I
3mm<r<7mm	II
7mm<r<10mm	III
r>10mm	IV

In this project, stages are classified as follows:

Stage 1 - The tumor is usually small and hasn’t grown outside of the organ.

Stage 2 & 3 - The tumor is larger or the cancerous tissues are becoming huge.

Stage 4 - The cancer has spread through the blood or lymphatic system to a distant site in the body (mthetastatic spread).

CHAPTER 3

MATLAB AND MACHINE LEARNING

3.1 MATLAB

The name MATLAB stands for MATrix LABoratory. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects.

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. Furthermore, MATLAB is a modern programming language environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research.

MATLAB has many advantages compared to conventional computer languages (e.g., C, FORTRAN) for solving technical problems. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. The software package has been commercially available since 1984 and is now considered as a standard tool at most universities and industries worldwide.

It has powerful built-in routines that enable a very wide variety of computations. It also has easy to use graphics commands that make the visualization of results immediately available. Specific applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering.

3.1.1 Creating MATLAB variables

MATLAB variables are created with an assignment statement. The syntax of variable assignment is

variable name = a value (or an expression)

For example,

```
 >> x = expression
```

where expression is a combination of numerical values, mathematical operators, variables, and function calls. On other words, expression can involve:

- manual entry
- built-in functions
- user-defined functions

3.1.2 Overwriting variable

Once a variable has been created, it can be reassigned. In addition, if you do not wish to see the intermediate results, you can suppress the numerical output by putting a semicolon (;) at the end of the line. Then the sequence of commands looks like this:

```
 >> t = 5;
>> t = t+1
 t = 6
```

3.1.3 Error messages

If we enter an expression incorrectly, MATLAB will return an error message. For example, in the following, we left out the multiplication sign, *, in the following expression

```
 >> x = 10;
 >> 5x ??? 5x
 |
 Error: Unexpected MATLAB expression.
```

3.1.4 Managing the workspace

The contents of the workspace persist between the executions of separate commands. Therefore, it is possible for the results of one problem to have an effect on the next one. To avoid this possibility, it is a good idea to issue a clear command at the start of each new independent calculation.

```
 >> clear
```

The command clear or clear all removes all variables from the workspace. This frees up system memory. In order to display a list of the variables currently in the memory, type

```
>> who
```

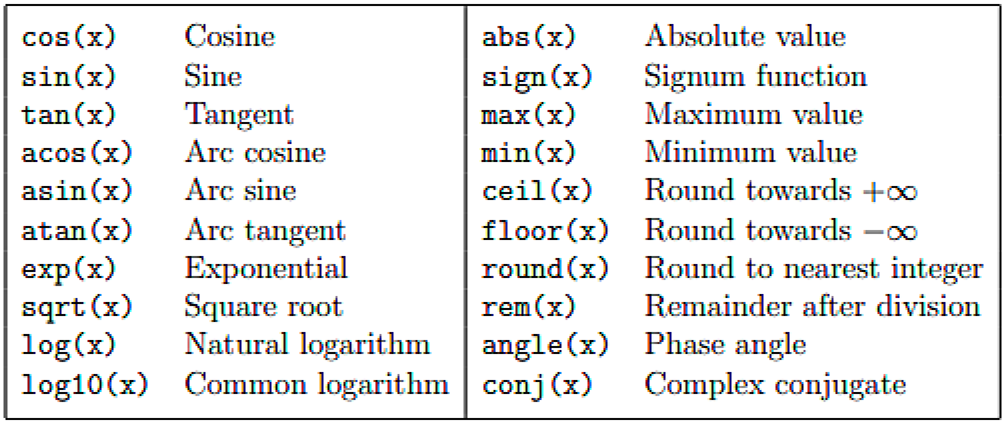
while, whos will give more details which include size, space allocation, and class of the variables.

3.1.5 Matlab functions

MATLAB offers many predefined mathematical functions for technical computing which contains a large set of mathematical functions.

Typing help elfun and help specfun calls up full lists of elementary and special functions respectively.

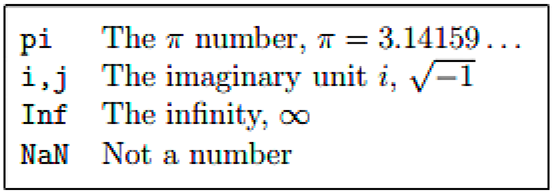
There is a long list of mathematical functions that are built into MATLAB. These functions are called built-ins. Many standard mathematical functions, such as sin(x), cos(x), tan(x), e x , ln(x), are evaluated by the functions sin, cos, tan, exp, and log respectively in MATLAB.



`cos(x)`	Cosine	`abs(x)`	Absolute value
`sin(x)`	Sine	`sign(x)`	Signum function
`tan(x)`	Tangent	`max(x)`	Maximum value
`acos(x)`	Arc cosine	`min(x)`	Minimum value
`asin(x)`	Arc sine	`ceil(x)`	Round towards $+\infty$
`atan(x)`	Arc tangent	`floor(x)`	Round towards $-\infty$
`exp(x)`	Exponential	`round(x)`	Round to nearest integer
`sqrt(x)`	Square root	`rem(x)`	Remainder after division
`log(x)`	Natural logarithm	`angle(x)`	Phase angle
`log10(x)`	Common logarithm	`conj(x)`	Complex conjugate

Figure 3.1 Elementary functions

In addition to the elementary functions, MATLAB includes a number of predefined 12 constant values



`pi`	The π number, $\pi = 3.14159\ldots$
`i,j`	The imaginary unit i, $\sqrt{-1}$
`Inf`	The infinity, ∞
`NaN`	Not a number

Figure 3.2 Constant Values

3.1.6 Creating simple plots

The basic MATLAB graphing procedure, for example in 2D, is to take a vector of xcoordinates, x = (x1, . . . , xN), and a vector of y-coordinates, y = (y1, . . . , yN), locate the points (xi , yi), with i = 1, 2, . . . , n and then join them by straight lines. You need to prepare x and y in an identical array form; namely, x and y are both row arrays or column arrays of the same length.

The MATLAB command to plot a graph is plot(x,y). The vectors x = (1, 2, 3, 4, 5, 6) and y = (3, −1, 2, 4, 5, 1).

```
>> x = [1 2 3 4 5 6];
>> y = [3 -1 2 4 5 1];
>> plot(x,y)
```

The plot functions has different forms depending on the input arguments. If y is a vector plot(y)produces a piecewise linear graph of the elements of y versus the index of the elements of y. If we specify two vectors, as mentioned above, plot(x,y) produces a graph of y versus x.

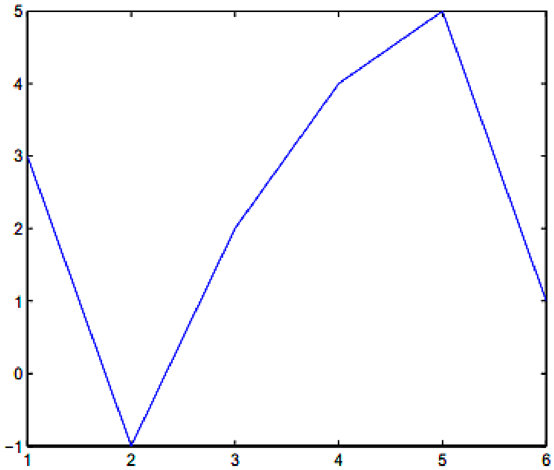


Figure 3.3 Plot for the vectors x and y

```
>> x = 0:pi/100:2*pi;
 >> y = sin(x);
 >> plot(x,y)
```

3.1.7 Adding titles, axis labels, and annotations

MATLAB enables you to add axis labels and titles. For example, using the graph from the previous example, add an x- and y-axis labels.

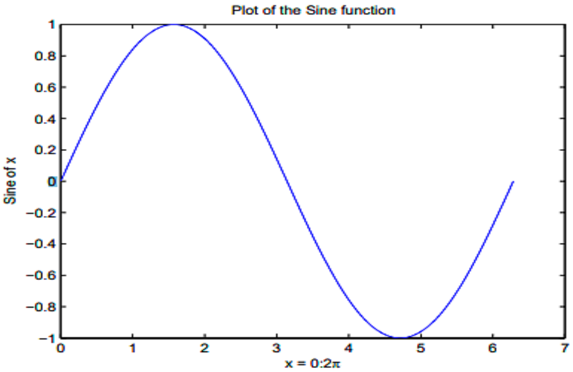


Figure 3.4 Plot of the Sine function

```
>> xlabel(’x = 0:2\pi’)
>> ylabel(’Sine of x’)
 >> title(’Plot of the Sine function’)
```

The color of a single curve is, by default, blue, but other colors are possible. The desired color is indicated by a third argument. For example, red is selected by plot(x,y,’r’). Note the single quotes, ’ ’, around r.

3.1.8 Multiple data sets in one plot

Multiple (x, y) pairs arguments create multiple graphs with a single call to plot. For example, these statements plot three related functions of x: y1 = 2 cos(x), y2 = cos(x), and y3 = 0.5 ∗ cos(x), in the interval 0 ≤ x ≤ 2π.

```
 >> x = 0:pi/100:2*pi;
 >> y1 = 2*cos(x);
 >> y2 = cos(x);
 >> y3 = 0.5*cos(x);
 >> plot(x,y1,’--’,x,y2,’-’,x,y3,’:’)
 >> xlabel(’0 \leq x \leq 2\pi’)
 >> ylabel(’Cosine functions’)
>> legend(’2*cos(x)’,’cos(x)’,’0.5*cos(x)’) 16
 >> title(’Typical example of multiple plots’)
 >> axis([0 2*pi -3 3])
```

The result of multiple data sets in one graph plot is shown in figure 3.5

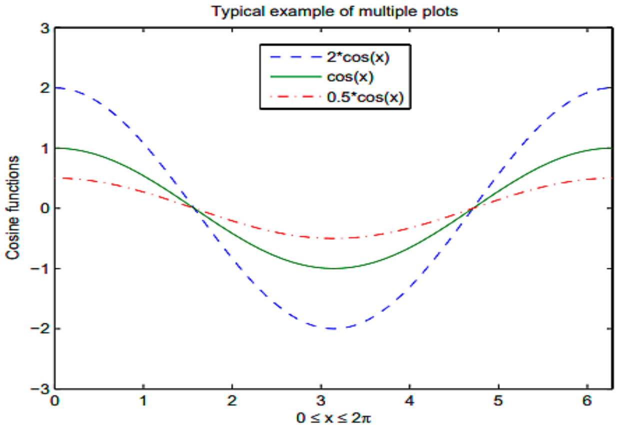


Figure 3.5 Typical example of multiple plot

3.2 MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. They will be required to help identify the most relevant business questions and the data to answer them. It works on

1. A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate about a pattern in the data.
2. An Error Function: An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.
3. A Model Optimization Process: If the model can fit better to the data points in the

training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this “evaluate and optimize” process, updating weights autonomously until a threshold of accuracy has been met.

3.2.1 Machine Learning Methods

Machine learning models fall into three primary categories.

3.2.1.1 Supervised machine learning

Supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, and support vector machine (SVM).

3.2.1.2 Unsupervised machine learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. This method’s ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It’s also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, and probabilistic clustering methods.

3.2.1.3 Semi-supervised learning

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it’s too costly to label enough data.

3.2.2 Common Machine Learning Algorithms

Numerous machine learning algorithms are commonly used. These include:

- **Neural networks:** Neural networks simulate the way the human brain works, with a huge number of linked processing nodes. Neural networks are good at recognizing patterns and play an important role in applications including natural language translation, image recognition, speech recognition, and image creation.
- **Linear regression:** This algorithm is used to predict numerical values, based on a linear relationship between different values. For example, the technique could be used to predict house prices based on historical data for the area.
- **Logistic regression:** This supervised learning algorithm makes predictions for categorical response variables, such as“yes/no” answers to questions. It can be used for applications such as classifying spam and quality control on a production line.
- **Clustering:** Using unsupervised learning, clustering algorithms can identify patterns in data so that it can be grouped. Computers can help data scientists by identifying differences between data items that humans have overlooked.
- **Decision trees:** Decision trees can be used for both predicting numerical values (regression) and classifying data into categories. Decision trees use a branching sequence of linked decisions that can be represented with a tree diagram. One of the advantages of decision trees is that they are easy to validate and audit, unlike the black box of the neural network.

- **Random forests:** In a random forest, the machine learning algorithm predicts a value or category by combining the results from a number of decision trees.

3.2.3 Real-World Machine Learning Use cases

Here are just a few examples of machine learning you might encounter every day:

- **Speech recognition:** It is also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, and it is a capability which uses natural language processing (NLP) to translate human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or improve accessibility for texting.
- **Customer service:** Customer service: Online chatbots are replacing human agents along the customer journey, changing the way we think about customer engagement across websites and social media platforms. Chatbots answer frequently asked questions (FAQs) about topics such as shipping, or provide personalized advice, cross-selling products or suggesting sizes for users. Examples include virtual agents on e-commerce sites; messaging bots, using Slack and Facebook Messenger; and tasks usually done by virtual assistants and voice assistants.
- **Computer vision:** This AI technology enables computers to derive meaningful information from digital images, videos, and other visual inputs, and then take the appropriate action. Powered by convolutional neural networks, computer vision has applications in photo tagging on social media, radiology imaging in healthcare, and self-driving cars in the automotive industry.
- **Recommendation engines:** Using past consumption behavior data, AI algorithms can help to discover data trends that can be used to develop more effective cross-selling strategies. This approach is used by online retailers to make relevant product recommendations to customers during the checkout process.
- **Automated stock trading:** Designed to optimize stock portfolios, AI-driven high-frequency trading platforms make thousands or even millions of trades per day without

human intervention.

- **Fraud detection:** Banks and other financial institutions can use machine learning to spot suspicious transactions. Supervised learning can train a model using information about known fraudulent transactions. Anomaly detection can identify transactions that look atypical and deserve further investigation.

3.2.4 Challenges of Machine Learning

As machine learning technology has developed, it has certainly made our lives easier. However, implementing machine learning in businesses has also raised a number of ethical concerns about AI technologies. Some of these include:

- **Technological singularity**

While this topic garners a lot of public attention, many researchers are not concerned with the idea of AI surpassing human intelligence in the near future. Technological singularity is also referred to as strong AI or superintelligence. Philosopher Nick Bostrum defines superintelligence as “any intellect that vastly outperforms the best human Brain in practically every field, including scientific creativity, general wisdom, and social skills.” Despite the fact that superintelligence is not imminent in society, the idea of it raises some interesting questions as we consider the use of autonomous systems, like self-driving cars. It’s unrealistic to think that a driverless car would never have an accident, but who is responsible and liable under those circumstances? Should we still develop autonomous vehicles, or do we limit this technology to semi-autonomous vehicles which help people drive safely? The jury is still out on this, but these are the types of ethical debates that are occurring as new, innovative AI technology develops.

AI impact on jobs

While a lot of public perception of artificial intelligence centers around job losses, this concern should probably be reframed. With every disruptive, new technology, we see that the market demand for specific job roles shifts. For example, when we look at the automotive industry, many manufacturers, like GM, are shifting to focus on electric vehicle production to align with green initiatives. The energy industry isn’t going away, but the source of energy is shifting from a fuel economy to an electric one.

In a similar way, artificial intelligence will shift the demand for jobs to other areas. There will need to be individuals to help manage AI systems. There will still need to be people to address more complex problems within the industries that are most likely to be affected by job demand shifts, such as customer service. The biggest challenge with artificial intelligence and its effect on the job market will be helping people to transition to new roles that are in demand.

- **Privacy**

Privacy tends to be discussed in the context of data privacy, data protection, and data security. These concerns have allowed policymakers to make more strides in recent years. For example, in 2016, GDPR legislation was created to protect the personal data of people in the European Union and European Economic Area, giving individuals more control of their data. In the United States, individual states are developing policies, such as the California Consumer Privacy Act (CCPA), which was introduced in 2018 and requires businesses to inform consumers about the collection of their data. Legislation such as this has forced companies to rethink how they store and use personally identifiable information (PII). As a result, investments in security have become an increasing priority for businesses as they seek to eliminate any vulnerabilities and opportunities for surveillance, hacking, and cyberattacks.

- **Bias and discrimination**

Instances of bias and discrimination across a number of machine learning systems have raised many ethical questions regarding the use of artificial intelligence. How can we safeguard against bias and discrimination when the training data itself may be generated by biased human processes? While companies typically have good intentions for their automation efforts, Reuters highlights some of the unforeseen consequences of incorporating AI into hiring practices. In their effort to automate and simplify a process, Amazon unintentionally discriminated against job candidates by gender for technical roles, and the company ultimately had to scrap the project. Harvard Business Review has raised other pointed questions about the use of AI in hiring practices, such as what data you should be able to use when evaluating a candidate for a role.

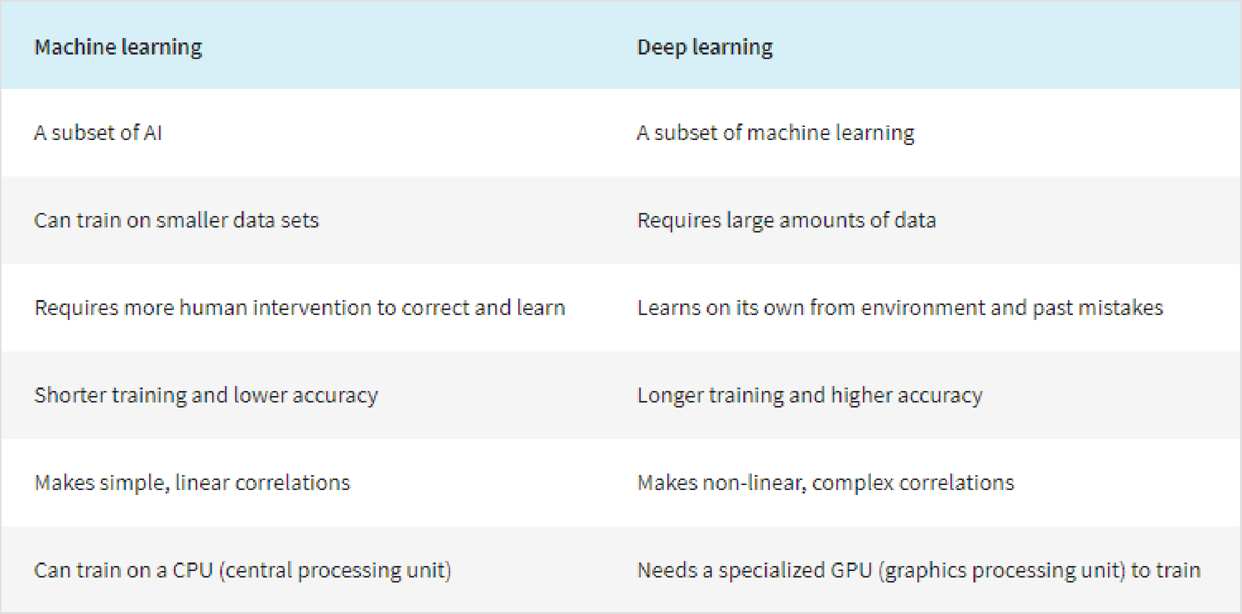
Bias and discrimination aren't limited to the human resources function either; they can be found in a number of applications from facial recognition software to social media algorithms.

As businesses become more aware of the risks with AI, they've also become more active in this discussion around AI ethics and values.

- **Accountability**

Since there isn't significant legislation to regulate AI practices, there is no real enforcement mechanism to ensure that ethical AI is practiced. The current incentives for companies to be ethical are the negative repercussions of an unethical AI system on the bottom line. To fill the gap, ethical frameworks have emerged as part of a collaboration between ethicists and researchers to govern the construction and distribution of AI models within society. However, at the moment, these only serve to guide. Some research shows that the combination of distributed responsibility and a lack of foresight into potential consequences aren't conducive to preventing harm to society.

3.2.5 Machine Learning vs Deep Learning



Machine learning	Deep learning
A subset of AI	A subset of machine learning
Can train on smaller data sets	Requires large amounts of data
Requires more human intervention to correct and learn	Learns on its own from environment and past mistakes
Shorter training and lower accuracy	Longer training and higher accuracy
Makes simple, linear correlations	Makes non-linear, complex correlations
Can train on a CPU (central processing unit)	Needs a specialized GPU (graphics processing unit) to train

Figure 3.6 Machine Learning vs Deep Learning

CHAPTER 4

Multiclass SVM

For the machine to be able to decide how to assign an instance to its group, it has to learn the patterns of that assignment from the training features available in a labeled training data set. There are two types of classification: binary classification and multiclass classification.

4.1 BINARY CLASSIFICATION

In this type, the machine should classify an instance as only one of two classes; yes/no, 1/0, or true/false.

The classification question in this type is always in the form of yes/no. For example, does this image contain a human? Does this text has a positive sentiment? Will the price of a particular stock increase in the next month?

4.2 MULTICLASS CLASSIFICATION

In this type, the machine should classify an instance as only one of three classes or more.

The following are examples of multiclass classification:

- Classifying a text as positive, negative, or neutral
- Determining the dog breed in an image
- Categorizing a news article to sports, politics, economics, or social

4.3 SUPPORT VECTOR MACHINES (SVM)

SVM is a supervised machine learning algorithm that helps in classification or regression problems. It aims to find an optimal boundary between the possible outputs.

In the base form, linear separation, SVM tries to find a line that maximizes the separation between a two-class data set of 2-dimensional space points. To generalize, the objective is to find a hyperplane that maximizes the separation of the data points to their potential classes in an -dimensional space. The data points with the minimum distance to the hyperplane (closest points) are called Support Vectors.

In the image below, the Support Vectors are the 3 points (2 blue and 1 green) laying on the scattered lines, and the separation hyperplane is the solid red line:

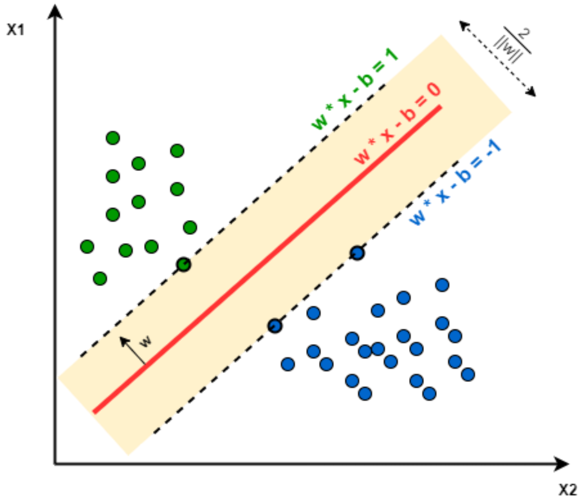


Figure 4.1 Support Vectors in SVM

The computations of data points separation depend on a kernel function. There are different kernel functions: Linear, Polynomial, Gaussian, Radial Basis Function (RBF), and Sigmoid. Simply put, these functions determine the smoothness and efficiency of class separation, and playing around with their hyperparameters may lead to overfitting or underfitting.

4.3.1 Multiclass Classification Using SVM

In its most simple type, SVM doesn’t support multiclass classification natively. It supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems.

The idea is to map data points to high dimensional space to gain mutual linear separation between every two classes. This is called a One-to-One approach, which breaks down the multiclass problem into multiple binary classification problems. A binary classifier per each pair of classes.

Another approach one can use is One-to-Rest. In that approach, the breakdown is set to a binary classifier per each class.

A single SVM does binary classification and can differentiate between two classes. So that, according to the two breakdown approaches, to classify data points from classes data set:

- In the One-to-Rest approach, the classifier can use m SVMs. Each SVM would predict membership in one of the classes.

- In the One-to-One approach, the classifier can use m(m-1)/2 SVMs.

Let’s take an example of 3 classes classification problem; green, red, and blue, as the following image:

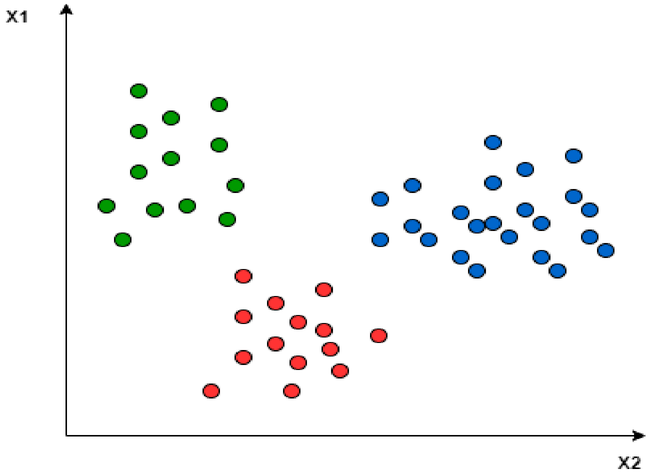


Figure 4.2 3 Classes classification

Applying the two approaches to this data set results in the followings:

In the One-to-One approach, we need a hyperplane to separate between every two classes, neglecting the points of the third class. This means the separation takes into account only the points of the two classes in the current split. For example, the red-blue line tries to maximize the separation only between blue and red points. It has nothing to do with green points:

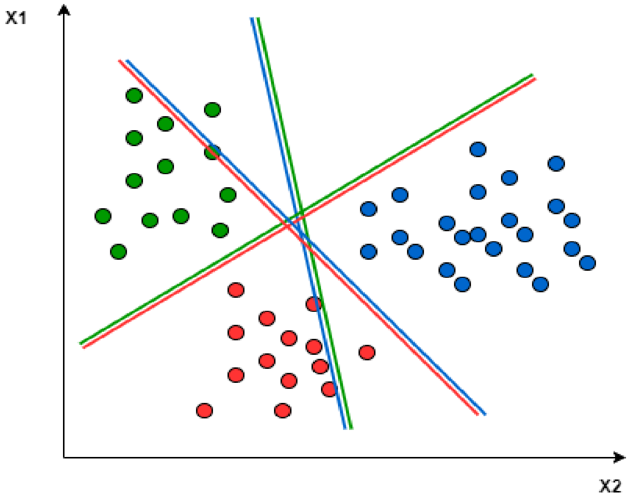


Figure 4.3 One-to-One Approach

In the One-to-Rest approach, we need a hyperplane to separate between a class and all others at once. This means the separation takes all points into account, dividing them into two groups; a group for the class points and a group for all other points. For example, the green line tries to maximize the separation between green points and all other points at once:

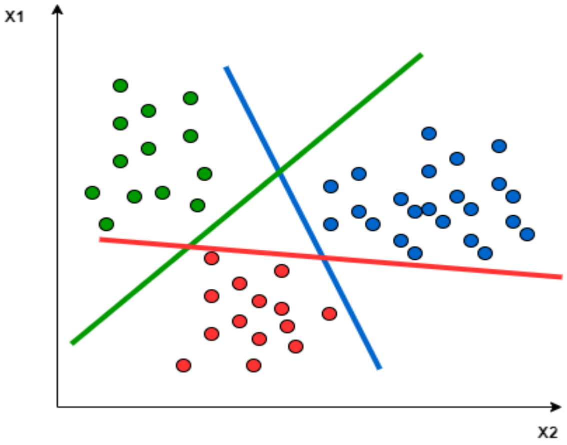


Figure 4.4 One-to-Rest Approach

One of the most common real-world problems for multiclass classification using SVM is text classification. For example, classifying news articles, tweets, or scientific papers.

4.3.2 Advantages of Support Vector Machine

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient

4.3.3 Disadvantages of Support Vector Machine

1. SVM algorithm is not suitable for large data sets.
2. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.

3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 RESULTS

A data set containing 2574 CT scan images of Brain were collected from Cancer Imaging Archive (CIA) data base. In detection of tumor regions, our algorithm has detected a total of 4906 tumor regions from all the 2574 images. We trained the SVM model with this training data set. On imposing a new image (Fig 5.1) of unknown stage was given as the input to the trained model to predict the stage of the, we extracted the outputs of the image at various processing steps (shown from Fig 5.1 to Fig 5.11). The results of the stage prediction by the machine learning algorithm is as demonstrated,

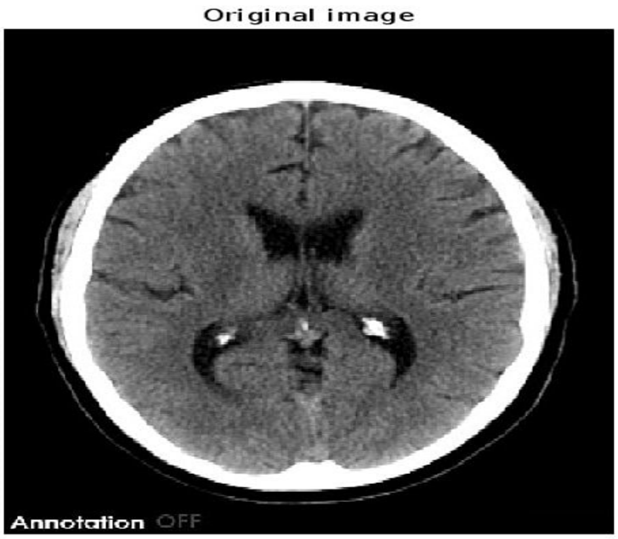


Figure 5.1 Original Image

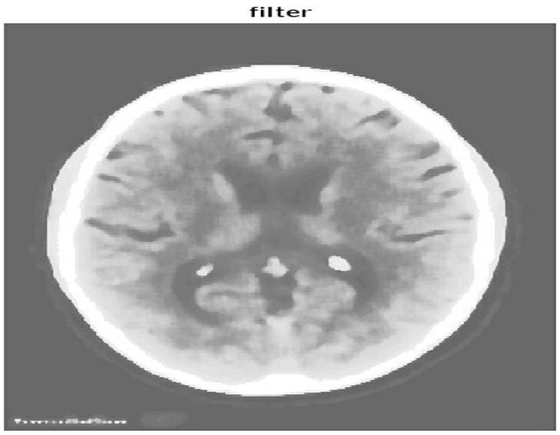


Figure 5.2 Filter

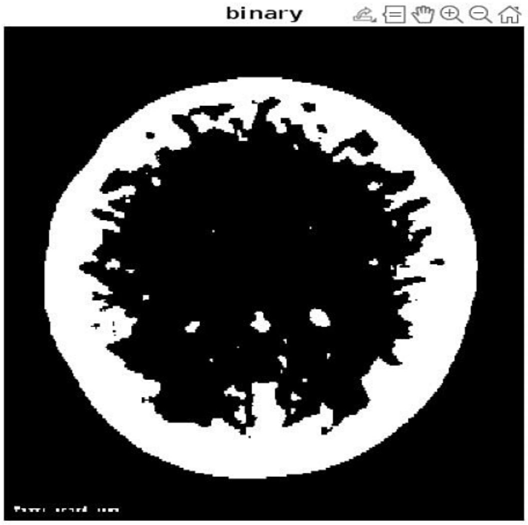


Figure 5.3 Binary Image

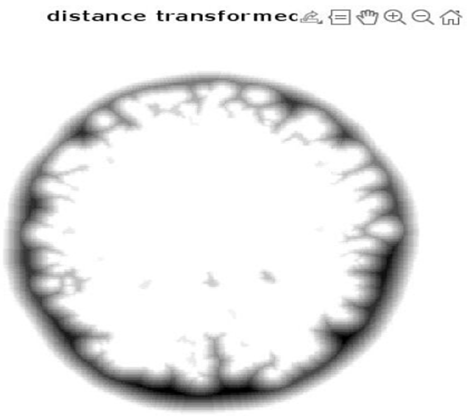


Figure 5.4 Distance Transform

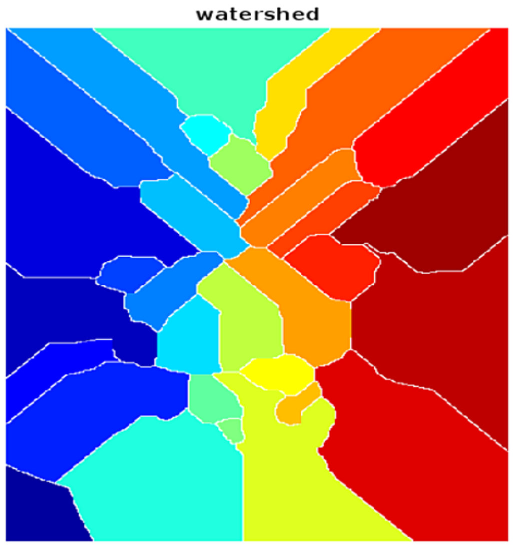


Figure 5.5 Watershed

We can observe that, the tumor regions are more accurately segmented with H-minima transform. The regions of the tumors from the segmentation step are

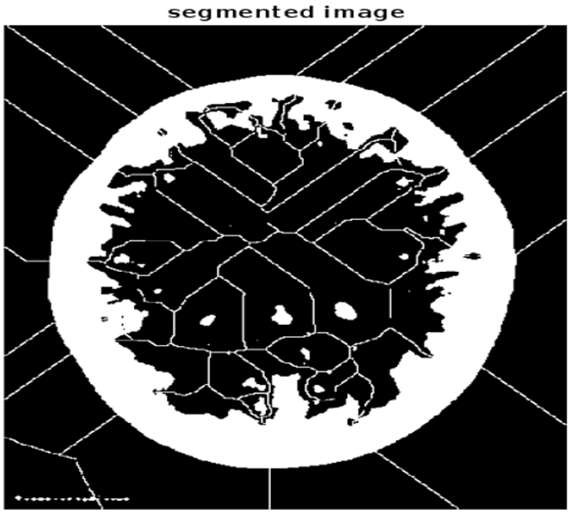


Figure 5.6 Segmented Image

The Region of Interest in detection of Brain cancer are tumor regions that can be obtained from Morphological operations on the watershed segmented image. We used erosion and opening operations on the watershed segmented image. Erosion operation is performed to remove the tiny regions which can easily identify the blood vessels and tumor regions. On erosion, blood vessels which are in small in size are easily removed. Opening operation is erosion followed by dilation. The tumor areas which are connected to Brain walls are preserved on opening the image. Subtracting the eroded image from the image on which the opening operation is performed results in tumor extraction. Finally, Tumor regions are extracted as the output from segmentation.

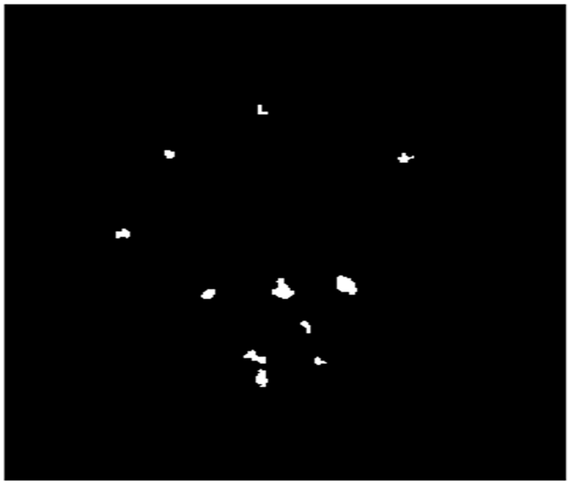


Figure 5.7 Tumor

The regions that were appeared in Figure 5.8 are not all cancer tumor regions. They can also be blood tissues too. The extracted tumors having greater than 55 pixels are only considered to be exact tumors. They were shown in Figure 5.9.

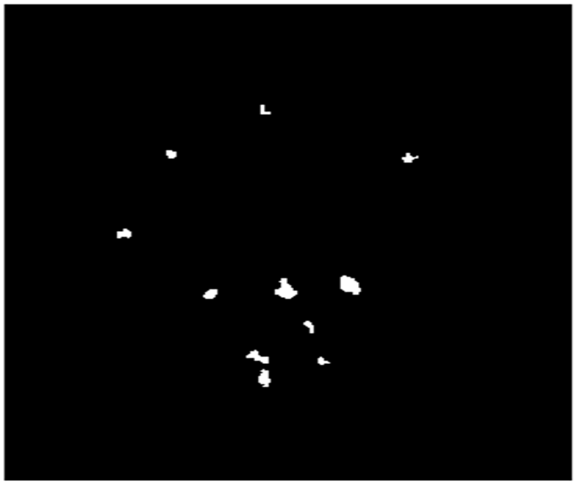


Figure 5.8 Exact Tumors

The tumor regions having eccentricity < 0.98 are only considered as Brain cancer tumors. So, the region of interest from the segmentation is shown in below figure

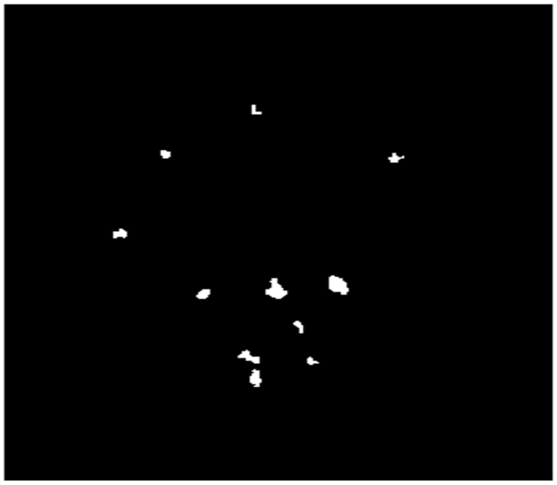


Figure 5.9 Tumor or Not

Boundary tracing followed by labelling the tumor regions is performed on the segmented output for better interpretation of the tumor regions. In Feature extraction, some of the important region-based features or parameters of the image like area in mm^2, perimeter in mm, centroid, eccentricity and orientation are extracted using region props. From the radius obtained, the stage of the cancer is estimated by the radiologists.

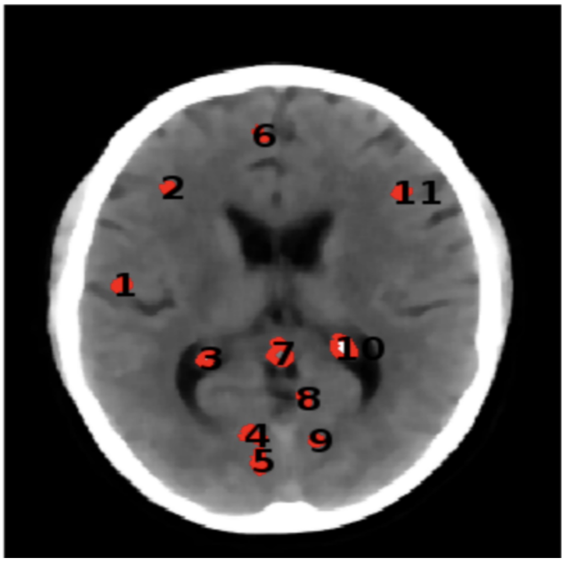


Figure 5.10 outlines

5.2 ADVANTAGES

There are several advantages to using machine learning and image processing for the detection of Brain cancer, including:

- **Early Detection:** Machine learning algorithms can detect subtle changes in images that may be indicative of early-stage Brain cancer. This can lead to early detection and treatment, which can improve the chances of survival.
- **Accuracy:** Machine learning algorithms can process large amounts of data quickly and accurately, which can reduce the risk of misdiagnosis and improve the accuracy of Brain cancer detection.
- **Efficiency:** Machine learning algorithms can analyze large volumes of medical images in a short period of time, which can help to reduce the workload of healthcare professionals and increase the efficiency of Brain cancer screening programs.
- **Personalized Treatment:** Machine learning algorithms can analyze patient data to develop personalized treatment plans based on the individual's unique characteristics and medical history. This can lead to more effective treatments and better outcomes for patients with Brain cancer.
- **Cost-Effective:** Machine learning algorithms can help to reduce the cost of Brain cancer screening and treatment by improving the accuracy of diagnosis, reducing the need for unnecessary tests and procedures, and enabling more efficient use of healthcare resources.

Overall, the use of machine learning and image processing for Brain cancer detection has the potential to improve the accuracy, efficiency, and effectiveness of screening and treatment programs, ultimately leading to better outcomes for patients.

5.3 DISADVANTAGES

While there are several advantages to using machine learning and image processing for Brain cancer detection, there are also some potential disadvantages that should be considered, including:

- **Lack of Human Expertise:** Machine learning algorithms can only identify patterns in images that they have been trained on. They may not be able to detect more subtle or complex features that a human expert could identify. Therefore, the lack of human expertise in interpreting the results of machine learning algorithms could lead to misdiagnosis or missed diagnoses.
- **Dependence on Data Quality:** Machine learning algorithms rely heavily on the quality and quantity of the data that they are trained on. If the data is incomplete, inaccurate, or biased, the algorithm may produce unreliable results.
- **Limited Availability of Data:** In some cases, there may be limited data available for training machine learning algorithms. This can make it difficult to develop accurate and effective algorithms for Brain cancer detection.
- **Ethical Concerns:** The use of machine learning algorithms for Brain cancer detection raises ethical concerns about the privacy and security of patient data, as well as the potential for algorithmic bias and discrimination.
- **Need for Validation:** Before machine learning algorithms can be used in clinical settings, they must be thoroughly validated and tested to ensure that they are accurate and reliable. This can be a time-consuming and resource-intensive process.

Overall, while machine learning and image processing have the potential to improve Brain cancer detection, it is important to consider these potential disadvantages and address them in the development and implementation of these technologies.

5.4 FUTURE SCOPE

This project has attempted to determine the stages of cancer, but the accuracy achieved is currently unsatisfactory. To address this issue, future research can explore alternative algorithms or approaches, such as utilizing deep learning algorithms, to achieve a more precise and accurate determination of cancer stages. By developing and implementing new methodologies, we can strive towards more accurate and reliable cancer diagnosis, which can potentially improve patient outcomes and overall healthcare.

5.5 CONCLUSION

Using MATLAB R2022a software, our algorithm has dealt with the drawback of watershed algorithm in segmentation called the over-segmentation and minimized it with H-minima transform and the results have shown that the SVM classifier has classified the stage of the tumor regions for a new input. This algorithm can be used by Radiologists to determine the stage of Brain cancer within no time so that the patient can be cured with necessary treatment as early as possible and can be prevented from further weakening due to Brain cancer.

REFERENCES

[1] Oneeza Tehreem Khan, D Rajeswari, “Brain Tumor detection Using Machine Learning and Deep Learning Approaches”, International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), vol. 21, Issue 1, pp. 123-156, April 2022.

[2] S.M. Moinuddin, "Brain Tumor Classification Using Deep Learning Based on Magnetic Resonance Imaging" IEEE Computer Science, Engineering and Applications, vol. 3, pp. 60-64, July2021 .

[3] A. Vaidya, ”Multi-Modal Brain Tumor Segmentation Using Convolutional Neural Networks", IEEE Inventive Research In Computing Applications, pp. 212-320, September 2021.

[4] Y. Huang, "Deep Learning-Based Radiomics for Brain Tumor Diagnosis and Prognosis Prediction Using Multimodal MRI Images" , IEEE Inventive Computational Technologies , vol. 3, no. 1, pp. 277-281, June 2021.

[5] Shardeep Kaur Sooch, Darpan Anand, Rajesh Deorari, "Brain Tumor detection with GLCM feature extraction and hybrid classification approach", 2022 International Conference on Cyber Resilience (ICCR), pp.1-5, 2022.