KIDNEY TUMOR SEGMENTATION AND CLASSIFICATION ON ABDOMINAL CT SCANS

Project – A Report

Submitted in complete fulfilment of requirements

For the degree of

Bachelor of Engineering

By

Bansari Dhiraj Shah

Charmi Dhiren Sawla

Shraddha Bhanushali

Supervisor:

Prof. Poonam Bhogle

(Department of Computer Engineering)

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

(2016-17)

Project Report Approval for B.E.

This project report entitled Kidney Tumor Segmentation and Classification on Abdominal scans by **Bansari Dhiraj Shah**, **Charmi Dhiren Sawla** and **Shraddha Bhanushali** is approved for the degree of Bachelor of Engineering.

Examiners

1.---

2.---

Supervisors

1.---

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Charmi .D. Sawla (Roll no. 1311109) ---------------------------------

(Signature)

Bansari .D. Shah (Roll no. 1311110) ---------------------------------

(Signature)

Shraddha .J. Bhanushali (Roll no. 1311070) ---------------------------------

(Signature)

Date: May 20 ,2017

Abstract:

Kidney Tumor consists of various kinds of cancer, where clear cell renal carcinoma accounts for the majority of cases. While many major risk factors linked with renal cancers have been identified, from hereditary illness, obesity, smoking and hypertension medication, at least 40% of these cases are of causes that have gone unexplained. Along with this, the accuracy of tumor detection screening varies with respect to image quality and expertise of the radiologist. Hence, to offset this variability it is highly desirable to develop algorithms that automatically detect the tumor regions and provide accurate segmented images along with an accurate cancer classification methodology. Keeping all this in mind, it is of utmost importance to correctly detect kidney tumor at early stage to prevent maximum damage. We have thus proposed a computer-assisted radiology system which will assess renal tumors in abdominal CT scans for the management of renal cell carcinoma diagnosis. In this study, we have implemented image segmentation using FCM algorithm, feature extraction using GLCM algorithm, and classification of tumor using SVM and KNN algorithm. The two classification algorithms and their respective results are presented and compared for their accuracy.

Acknowledgements:

The completion of our project has valuable contribution from various people with whom we interacted within the organization. This study, which is a complete fulfillment of the requirement for the degree of Bachelor of Engineering, would not have been possible without the constant support, mutual co-ordination of several people to whom we owe our sincere gratitude.

Firstly and most importantly, we would sincerely like to thank our mentor Prof. Poonam Bhogle. She has rendered her valuable knowledge of this subject and dedicated guidance with a touch of inspiration, vision and motivation. She has assisted us through any hurdles that we encountered by us giving plenty of early ideas, suggestions and modifications and encouraged us to complete the project to the best of our abilities.

We would like to thank our HOD Prof. Bharathi H.N. who extended every facility to us for making and completing this project smoothly. We would also like to thank our fellow peers and friends who constantly challenge us while also inspiring us, and provide us with their selfless help and invaluable advices. Lastly we thank the teaching and non – teaching staff of our college as well as the library staff for providing us timely essential information in the form of books and for their assistance in the laboratories.

Contents

Index

Figure Index

Table Index

Chapter 1. Introduction

1.1 Introduction:

Human body consists of myriad number of cells. When cell growth becomes uncontrollable the extra mass of cell transforms into tumor. CT scans and MRI are used for identification of tumor. The goal of our study is to accurately detect tumor and classify it through the means of several techniques involving medical image processing, pattern analysis, and computer vision for enhancement, segmentation and classification of kidney diagnosis. This system can be used by radiologists and healthcare specialists. The system is expected to improve the sensitivity, specificity, and diagnostic efficiency of kidney tumor screening using industry standard simulation software tool, MATLAB. These techniques involve pre-processing of CT scans collected from online cancer imaging archives as well as scans obtained from several pathology labs. Images are resized and then we apply the proposed algorithms for segmentation and classification. Computer-aided renal tumor system can be used by radiologists and health-care specialists. The system is expected to improve renal cancer screening procedure currently at use, and possibly reduce health care costs by decreasing the need for follow-up procedures such as biopsy. Several processing steps are required for the accurate characterization and analysis of biomedical image data.

1.2 Problem Definition

Our study deals with automated kidney tumor segmentation and classification. Normally the anatomy of kidney can be viewed by MRI scan or CT scan for diagnosis. Kidney tumor usually effects urinary bladder making detection of tumor is crucial for its treatment. It is important to predict the tumor and classify it so that appropriate treatment can be planned at an early stage. Different types of algorithms were developed for kidney tumor detection but they have some drawback in effective detection and extraction of tumor. Our System deals with implementation of simple and efficient algorithms for segmentation of kidney-tumor and also classification of tumor whether it is cancerous or non- cancerous. Tumor segmentation is done by fuzzy c-means algorithm, extraction of features is implemented by grey level co-occurrence matrix method (GLCM) and finally classification of tumor if it is benign or malignant is obtained by support vector machine tool (SVM) and k nearest neighbour classifier (KNN).

1.3 Scope

Our aim is to develop a semi-automated system for enhancement, segmentation and classification for kidney diagnosis. The system can be used by radiologists and healthcare specialists. The system incorporates image processing, pattern analysis, and computer vision techniques and is expected to improve the sensitivity, specificity, and efficiency of kidney tumor screening. The proper combination and parameterization of above phases enables the development of adjunct tools that can help on the early diagnosis or the monitoring of the therapeutic procedures.

Chapter 2. Literature Review

This shall normally present a critical appraisal of the previous work published in the literature pertaining to the topic of the investigation. The extent and emphasis of the chapter shall depend on the nature of the investigation. After performing the literature survey and analyzing various existing systems, we learnt about their advantages and drawbacks. Keeping those in mind, we have proposed a system which would minimize these drawbacks and even improve the efficiency of tumor screening.

Image processing is any form, of information processing, in which the input is an image. The existing method is based on the threshold and region growing[11]. At the threshold based segmentation the image is considered as having only two values either black or white. But the bitmap image contains 0 to 255 gray scale values. So it ignores the tumor cells also. In case of the region growing based segmentation it needs more user interaction for the selection of the seed. Seed is nothing but the center of the tumor cells; it may cause intensity inhomogeneity problems. And also it will not provide the acceptable result for all the images. The region growing method ignored the spatial characteristics. Normally spatial characteristics are important for malignant tumor detection. In thresholding. This is the main problem of the current system, due to that g based segmentation the image is considered as having only two values either black or white proposed technique for kidney tumor segmentation.

For many years, one of the most primary tasks in medical image processing has been to perform automatic image segmentation. A lot of methods have thus been developed and implemented from edge detection and following to region growing, region modelling and finally separation and mathematical morphology, etc. However, in some fields of application, such as medical or biomedical imaging, objects of interest (OOIs) aren’t well defined and even in the case of sophisticated automatic systems, it has been observed that they often fail. For such cases, the only solution until recently was found out to replace automatic methods by interactive or rather manual ones, wherein communication between the user and the imaging system is required, This type of interaction is quite tedious and there are high possibilities of fatigue errors since every stage of the algorithm is performed manually . Few years back, a third possibility was suggested by several researchers in the field of medical image processing: it consists in renouncing complete automatic image segmentation and classification in support of the usual semi-automatic methods, with minimal user interaction, in contrast to highly interactive approaches. In this study, we have thus tried to implement a semi-automated segmentation method in which the first stage of segmentation is done manually, and rest of the calculations related to feature extraction as well as classification processes are performed automatically.

Chapter 3. Project Management Plan

3.1 Feasibility Analysis

Economic feasibility: Whether the firm can afford to build the software, whether its benefits should substantially exceed its cost. Our project is economically feasible. Our system uses academic version of MATLAB R2016a; which was very feasible, economically since it can be viewed as a onetime investment.

Technical feasibility: Whether the technology needed for the system exists, how difficult it is to build. Our project is technically versatile system which can work on most platforms making it technically feasible to build requiring only few specifications. Software used for the project implementation is MATLAB. Basic technical knowledge of operating MATLAB software along with the classification toolbox is required for the developers.

Schedule Feasibility: How much time is available to build the new system, when it can be built. The project is entirely build from scratch to completion in a span of eight-nine months.

Ecological Feasibility: Whether the system has an impact on its environment. There are no adverse effects on the environment.

Operational feasibility: The system is easy to use and user-friendly. All maintenance issues will be handled efficiently. System is adaptable to most environments. Hence our system is operationally feasible.

3.2 Lifecycle Model

Waterfall model is non-iterative design process where System requirements are known initially and final outcome is determined .It progresses steadily downwards through above given faces.

When to use the waterfall model:

- This model is used only when the requirements are very well known, clear and fixed.
- Product definition is stable.
- Technology is understood.
- There are no ambiguous requirements
- Ample resources with required expertise are available freely

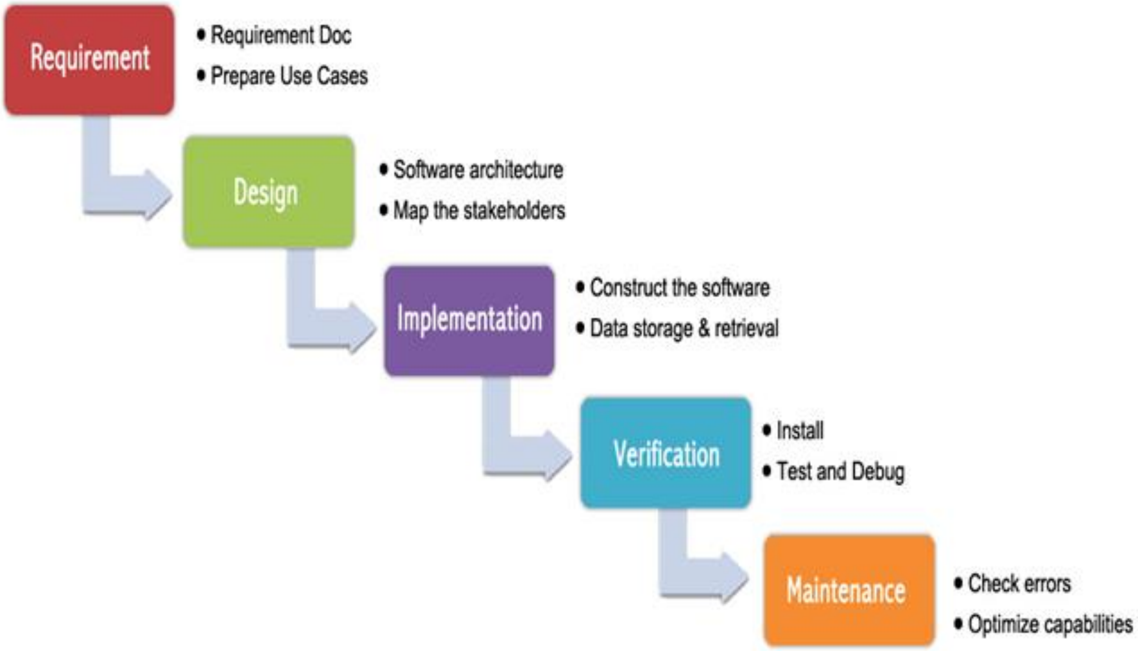


Fig 3.1 Waterfall Model

Functionality 1: Requirements Gathering

- MATLAB 2016a software and abdominal CT scan datasets.
- At least 100 datasets of abdominal scans. The vaster a dataset, more accurate will the results be.
- Gathering information about different filters such as-Gaussian ,Mean ,Gabor filters etc for seeing which gives the best result for our study
- Clustering algorithms for segmentation- Fuzzy-c means ,K means, graphing etc Feature extraction techniques and Classifiers such as SVM and ANN

Functionality 2: Design

Designing the process overview from applying filters to segmentation, feature extraction and classification of kidney tumor.

Functionality 3: Implementation

Implementing all algorithms in MATLAB.

Functionality 4: Verification

Verifying it by testing it on minimum 30 datasets of scans.

Functionality 5: Maintenance·

Maintaining from time to time for its efficiency.

Project deliverables

- Software Project Management Plan
- Software Requirements specifications
- Software Design Description
- System Test Document
- User Interface module Prototype
- User Manual
- Final Product

3.3 Project Cost and time estimation

Project only used the academic version of the MATLAB R2016. And no other additional cost was incurred. The system was operational by 15st March, 2017.

3.4 Resource plan

Resources:

There are 3 members working on this project – Charmi Sawla, Bansari Shah and Shraddha Bhanishali. Each of the team members will be involved in almost all the activities and will play a responsible role in the development of the prototype. The model being complex, it will include participation of every member of the team.

The project is guided by Prof. Poonam Bhogle who provided necessary assistance and guided all the team members. She analysed all the documents and reviewed them by submitting her opinions which are taken into due consideration.

Activities:

1. Write and submit proposal
2. Conduct literature survey
3. Data set collection.
4. Segmentation, feature extraction and Classification algorithms
5. Accuracy testing.
6. Write Software requirement specification(SRS)
7. Write Software project management plan
8. Write Software design description
9. Extensive accuracy testing
10. Modifications
11. Develop GUI.

Table 3.1**: Activities**

	Activity	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
	Resource											
April	**B**	**25%**	**25%**	**50%**								
	C	**25%**	**25%**	**50%**								

	S	25 %	25 %									
June	S		25 %									
	C		25 %									
	B		25 %									
July	S				25 %							
	C				25 %							
	B				25 %							
August	B				25 %							
	C				50 %							
	S				25 %							
September	C					20 %						
	B					50 %						
	S				50 %	30 %						
November	B						50 %					
	C							50 %				

	S						50%					
December	S								40%	25%		
	C									25%		
	B								20%	50%		
January	C									20%	50%	
	B									20%		
	S									30%	50%	
February	S										50%	
	B											
	C											50%
March	B											30%
	C											40%
	S											30%

Table 3.2: Activities conducted by resources every month

3.5 Task & Responsibility Assignment Matrix

Person	Responsibility	Task
Charmi Sawla	● Data Gathering ● Implementation ● Training ● Documentation ● Testing .	● Study of various IEEE papers, research on the previous projects and implementation techniques to be used. ● Data Collection and Analysis. ● Write Software requirement specification(SRS) ● Developing UML diagrams. ● Implementation of Functionalities for processing, and portfolio management. ● Generating timely reports and activities throughout our project

Bansari Shah	● Data Gathering ● Implementation ● Training ● Documentation ● Testing	● Study of various IEEE papers, research on the previous projects and implementation techniques to be used. ● Data Collection and Analysis. ● Write and submit proposal ● Conduct literature survey ● Designing GUI through which users will interact, designing simple functional model of project. ● Implementation of Functionalities for processing, and portfolio management. ● Testing each and every module for Debugging bugs and errors ● Generating timely reports and activities throughout our project

Shraddha .B	● Implementation ● Training ● Testing	● Study of various IEEE papers, research on the previous projects and implementation techniques to be used. ● Designing GUI through which users will interact, designing simple functional model of project. ● Implementation of Functionalities for processing, and portfolio management. ● Testing each and every module for Debugging bugs and errors ● Generating timely reports and activities throughout our project

Table 3.3: Task and responsibility matrix

3.6 Project Timeline Chart

Task	Start Date	End Date	Duration
Proposal	3rd March	15th March	12 days
Literature Survey	10th March	9th June	89 days
Fuzzy-C Means Algorithm	9th Aug	6th October	57 days
GLCM	16th October	26th October	20 days
KNN	30th October	20th Jan	50 days
SVM	30th Jan	10th Feb	11 days
Extensive accuracy testing .	1st Feb	15th Feb	14 days
SRS	4th March	8th March	4 days
Software Project Management Plan	24th Dec	30th Dec	6 days
Software Design Description	2nd Feb	8th Feb	6 days
Report	9th Dec	12th Dec	3 days
Extensive accuracy testing on more abdominal scans.	15th Feb	15th March	30 days
Modifications	16th March	13th April	27 days
GUI	30th Jan	10th Feb	45 days

Table 3.4: Project Timeline chart

Gantt Chart:

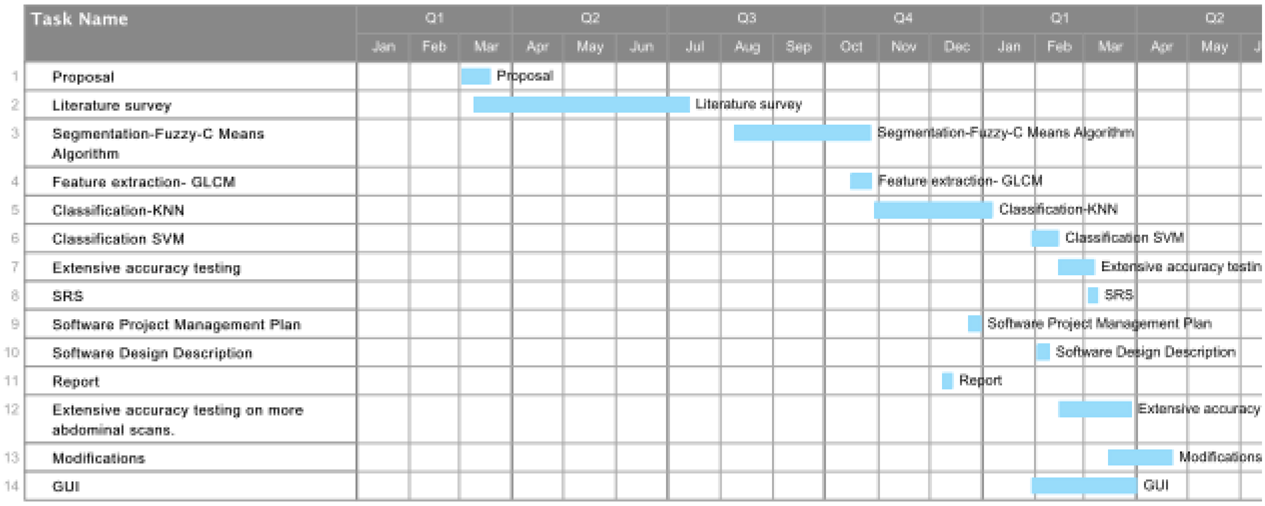


Fig. 3.2 Gantt Chart

Chapter 4. Project Analysis and Design

4.1 Software Architecture diagram

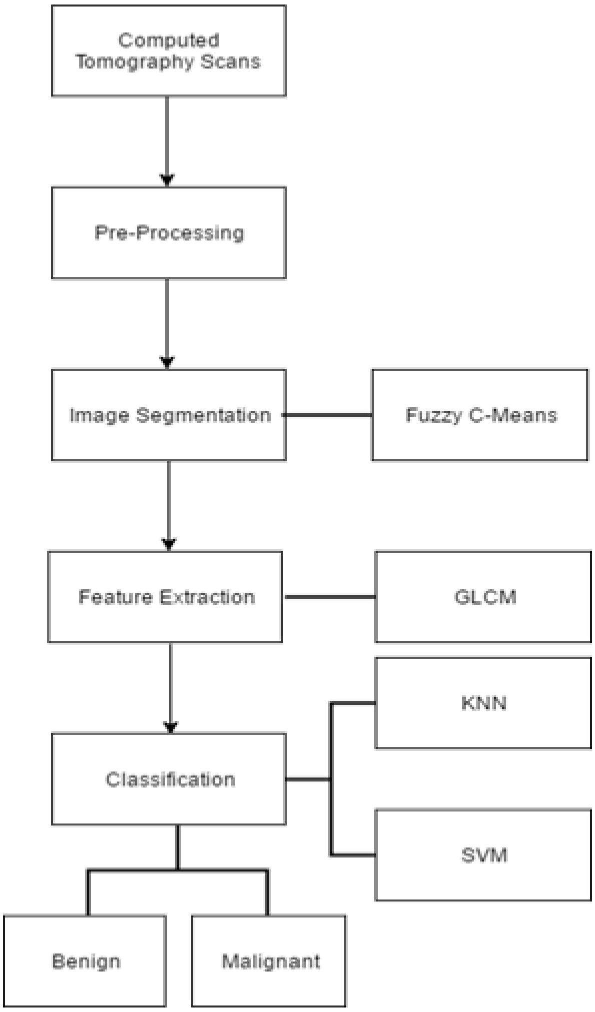


Fig 4.1 Flowchart and Software Architecture Diagram

4.2 Architectural style and justification

The diagram shows the various stages in the development of our system. The diagram shows interaction between the various components of the application and their position in the development hierarchy.

This style hence, is appropriate for the selected problem because all the modules in the selected problem function independently. The communication is strictly through message passing connectors. The flow of the system is from the top to the bottom.

4.3 Software Requirements Specification Document

Introduction

1. **Product overview**

Our objective is to develop a system incorporating image processing, and computer vision techniques for enhancement, segmentation of breast tumors. Our study aims at enhancing the current accuracy (diagnostic) of digital CT scans using industry standard simulation software tool, MATLAB and the online dataset.

These techniques involve pre-processing of digital CT scans by resizing them and then apply the proposed algorithms for segmentation, feature extraction and classification. The system is expected to improve the efficiency, sensitivity and specificity of renal cell carcinoma for kidney tumor screening, and possibly reduce health care costs by decreasing the need for follow-up procedures such as biopsy.

2. Specific requirements

2.1 External Interface Requirements

2.2 User Interfaces

- The end users of our system will be pathologists or radiologists.
- The process time delay between selecting a Tumor image and producing an output when using a user interface should be minimal.

2.3 Software product features

- The software is aimed to be simplistic, with minimal complexity and provides ease of use for even beginner users. Minimal training, if at all will be required to use the software product being built, which is provided by referring the user manuals and following the GUI diagrams and familiarising them with it.
- Rigorous training as well as testing of the dataset by a technical expert, in order to meet the required deliverables and have maximum throughput and efficiency for our system.
- Automation to ease the user of manual tasks is the primary aim of the project.

2.4 Software system attributes

· **Reliability:**

- Optimal functionality of software for as long as it is installed on the device.
- Minimal maintenance of system that can be automated as well.
- Must be adaptable to changing hardware and operating environment, making it versatile over time

- End user should be able to completely rely on the efficiency of the software.

· **Availability:**

- The software should be available on any platform to the user whenever required.
- In case of failure of software, consistency factor must be preserved.

· **Portability:**

- The software should be able to be run on 64 bit Windows operating system.
- It should be able to work seamlessly on any version of Windows from Windows 7 onwards on system and hardware any operating providing necessary portability

· **Maintainability:**

- Time to time maintenance of software, along with necessary updates is of utmost importance. This is especially needed if or when the error between predicted tumor and the actual tumor tends to vary much which may affect the efficiency of product. In such cases, the dataset is used for classification algorithms in the software to train to achieve certain level of accuracy.

· **Performance:**

- A good data storage repository having large capacity will be required to store the software as well as the CT scan datasets stored as input, as well as data collected from the inferences made as final output.

· **Security:**

- Administrator of the software must be given security so that the internal structure of the software stays unaltered by any non-authorized entity which may cause potential harm to the system.
- Access rights must be clearly and explicitly specified and practiced lawfully.
- Any kind of line-crossing by unauthorized personnel must be immediately restricted for minimal security damage.

2.5 Database requirements

- The database will contain the predicted data and the actual perceived data, as a percept sequence.
- The database consist of abdominal CT scans taken as input data, that are loaded to the system for training and classification, which are updated itself every time the patient database increases.
- Only the physician and radiologists will have access rights to modify this database so that sensitive information does not leak.

2.6 Performance Requirements

- To avoid any latency in operation or lag in performance of the application system, it is necessary to ensure processor is not slow, or that an outdated version of MATLAB is under use currently.

2.7 Safety Requirements

- User should not share dataset of patient with others.
- Only the physician and radiologists will have access rights to modify this database so that sensitive information does not leak.
- User should select the region of affected kidney out of the abdominal CT scans to provide for accurate mapping
- User should answer executive call when needed.

2.8 Security Requirements

- User and administrator should not share dataset of patient with another user.
- Only the physician and radiologists will have access rights to modify this database so that sensitive information does not leak.

2.9 Hardware and Software Platform requirements

Software Requirements:

- MATLAB 2016a
- Dataset obtained from pathology and radiology labs
- Online dataset from www.cancerimagingarchive.net
- Microsoft Office

Hardware Requirements:

- Intel Dual Core Dual Processor or advanced version
- Minimum 8GB of RAM
- Minimum 1 GB of Hard disk Space

4.4 Software Design Document

4.4.1 Use Case Diagram

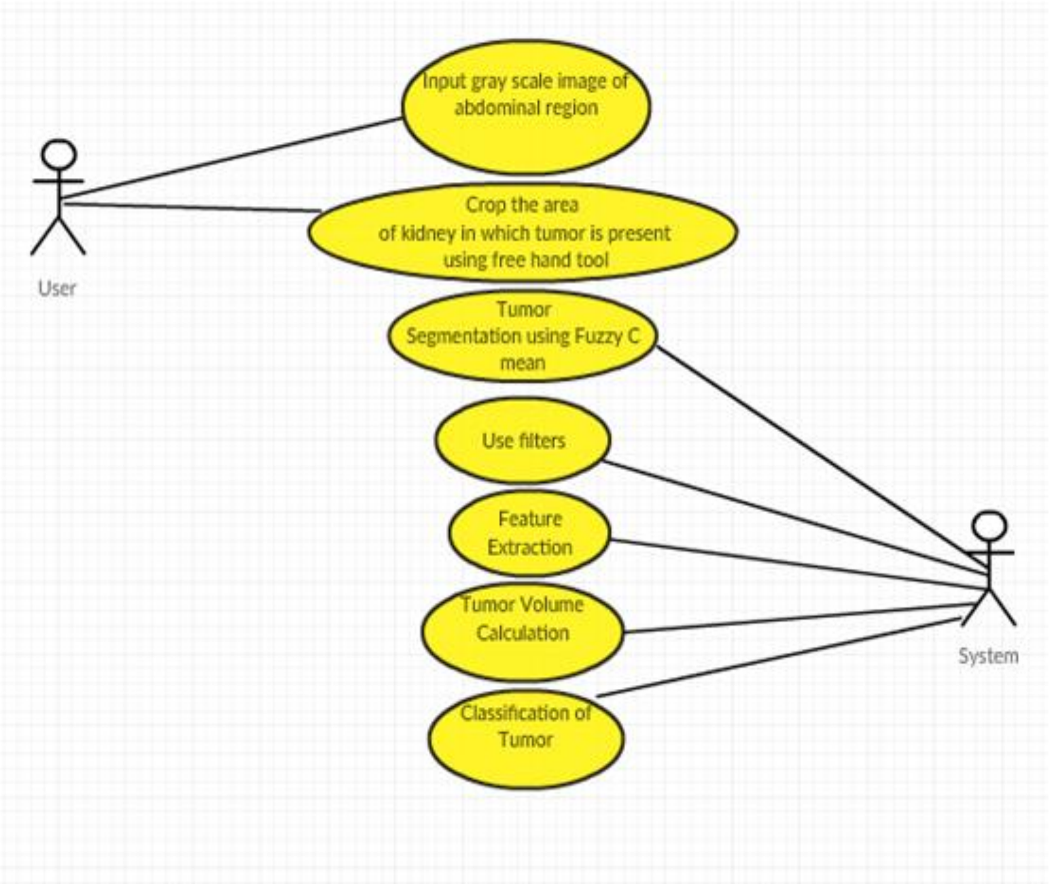

Input gray scale image of abdominal region
Crop the area of kidney in which tumor is present using free hand tool
Tumor Segmentation using Fuzzy C mean
Use filters
Feature Extraction
Tumor Volume Calculation
Classification of Tumor
User
System

Fig 4.2 Use Case Diagram

- The use case diagram consists of two actors, who interact with the software.
- The User: The user takes in input image and performs manual segmentation
- The System: The System performs all clustering, feature extraction, classification and training algorithms.

4.4.2 Class Diagram:

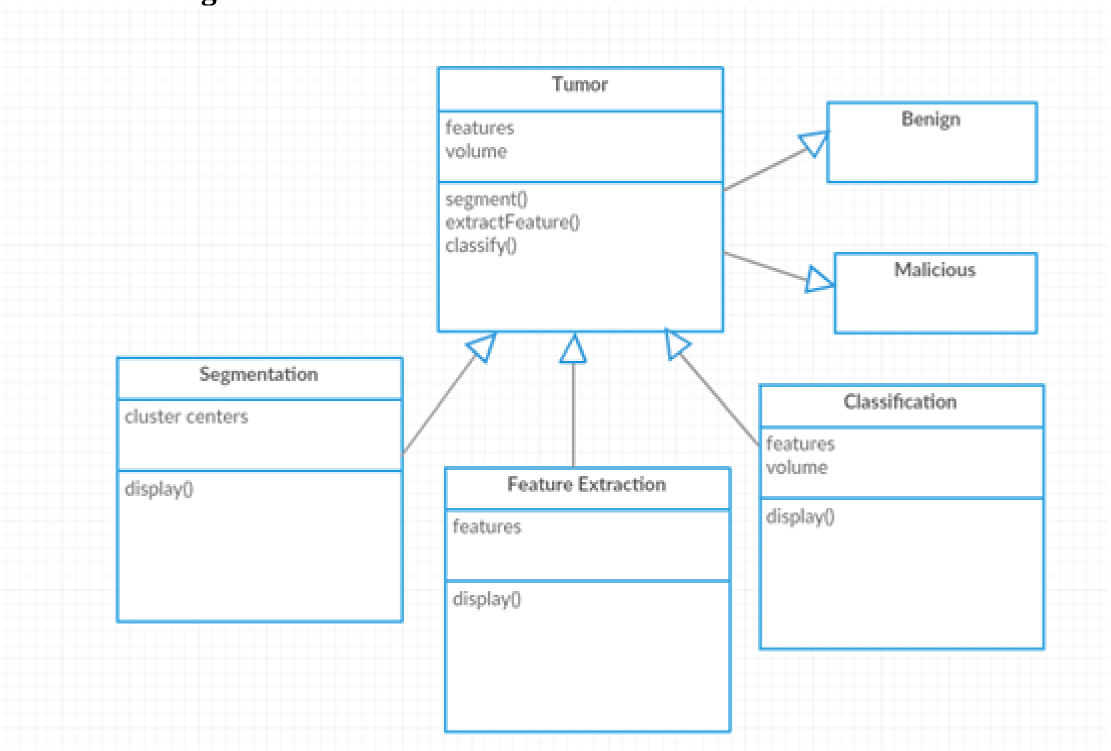


Fig 4.3 Class Diagram

4.4.3 UI:

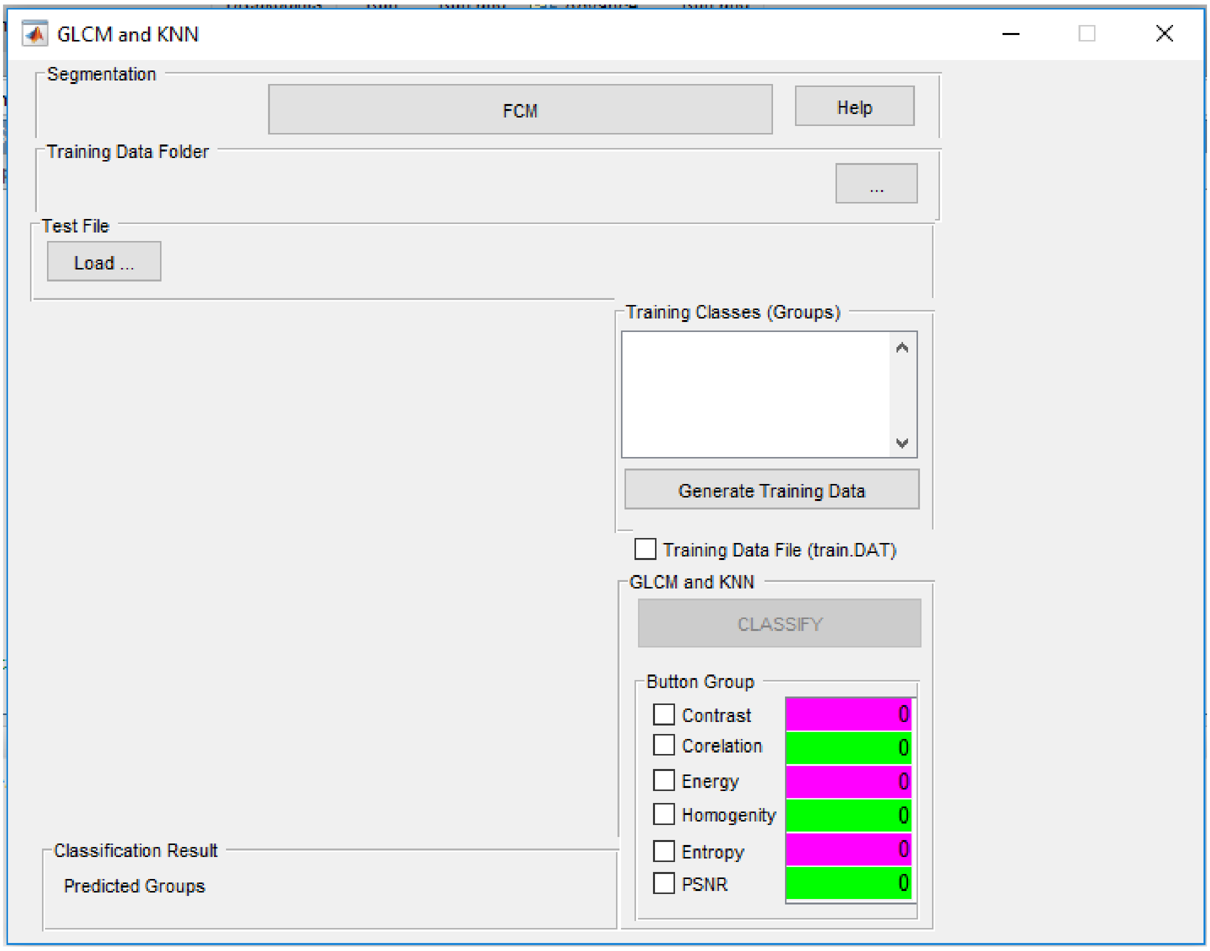


Fig 4.4 Simple UI

FCM

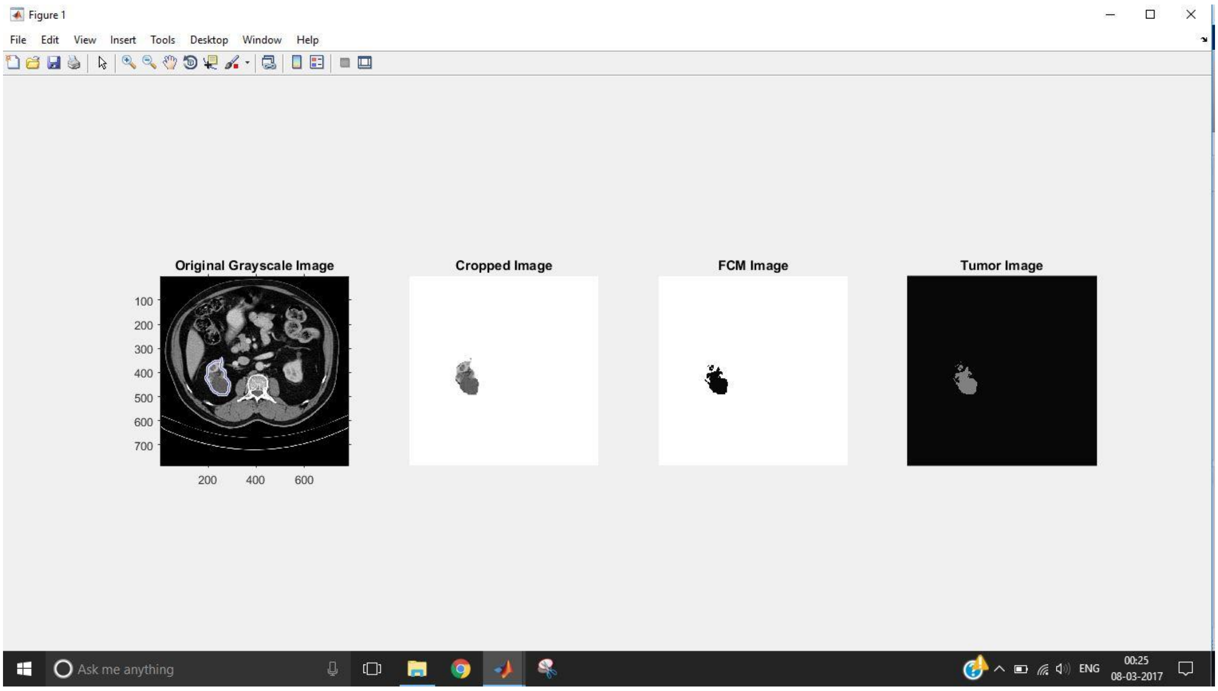
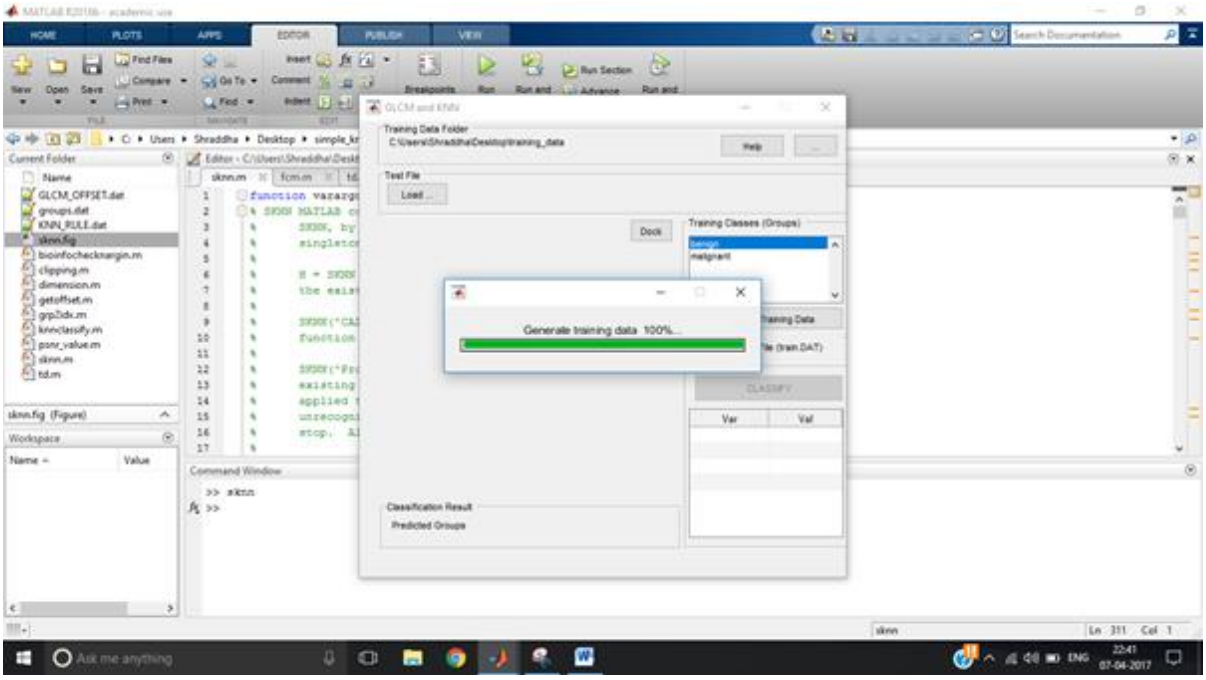


Fig 4.5 FCM Output

Training under process



Training Data Folder: C:\Users\Shraddha\Desktop\training_data

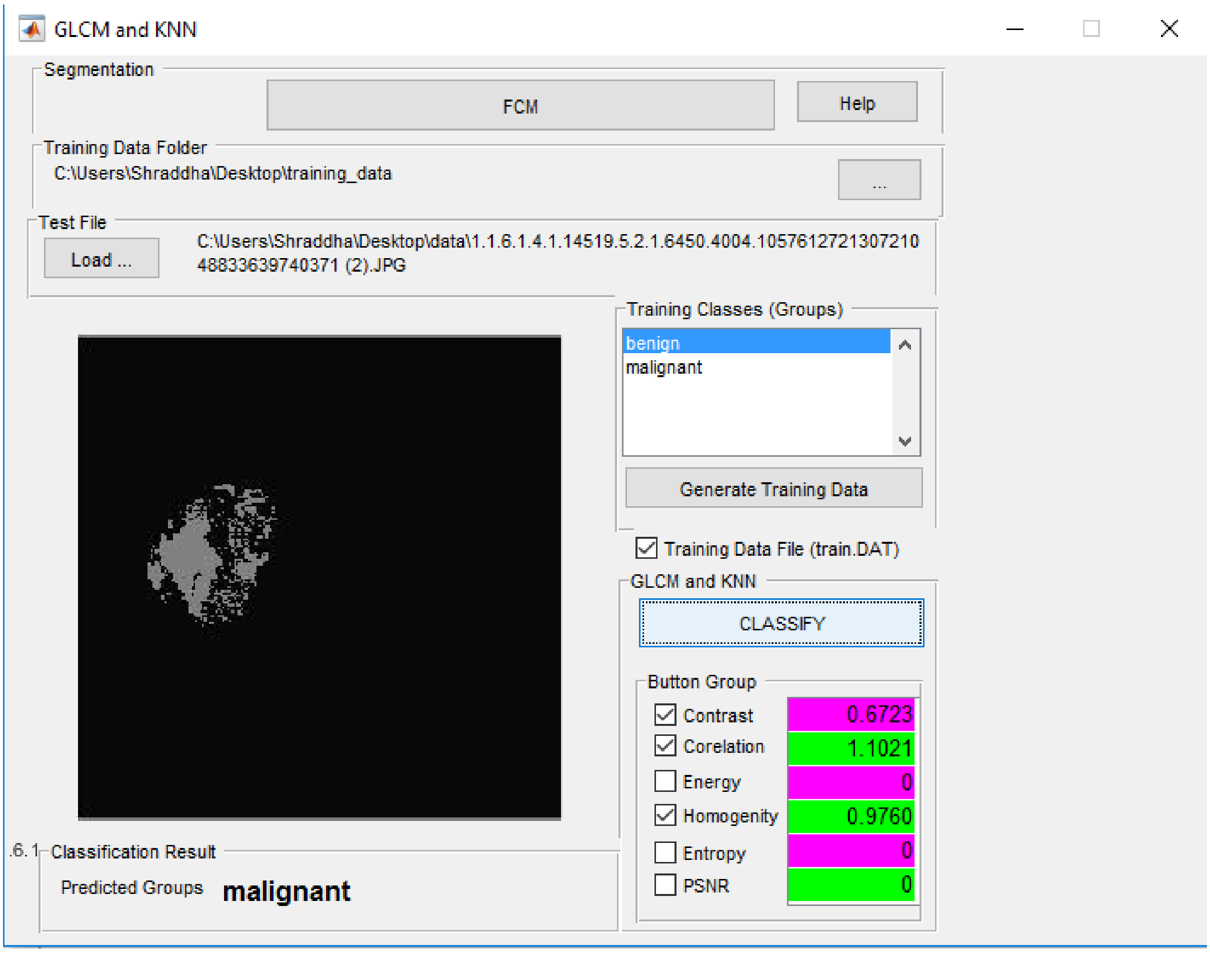
Training Classes (Groups): benign, malignant

Generate training data 100%...

```
>> sknn
>>
```

Fig 4.6 Training output

KNN Classification Displayed



GLCM and KNN

Segmentation

Training Data Folder
C:\Users\Shraddha\Desktop\training_data

Test File
C:\Users\Shraddha\Desktop\data\1.1.6.1.4.1.14519.5.2.1.6450.4004.105761272130721048833639740371 (2).JPG

Training Classes (Groups)
benign
malignant

☑ Training Data File (train.DAT)

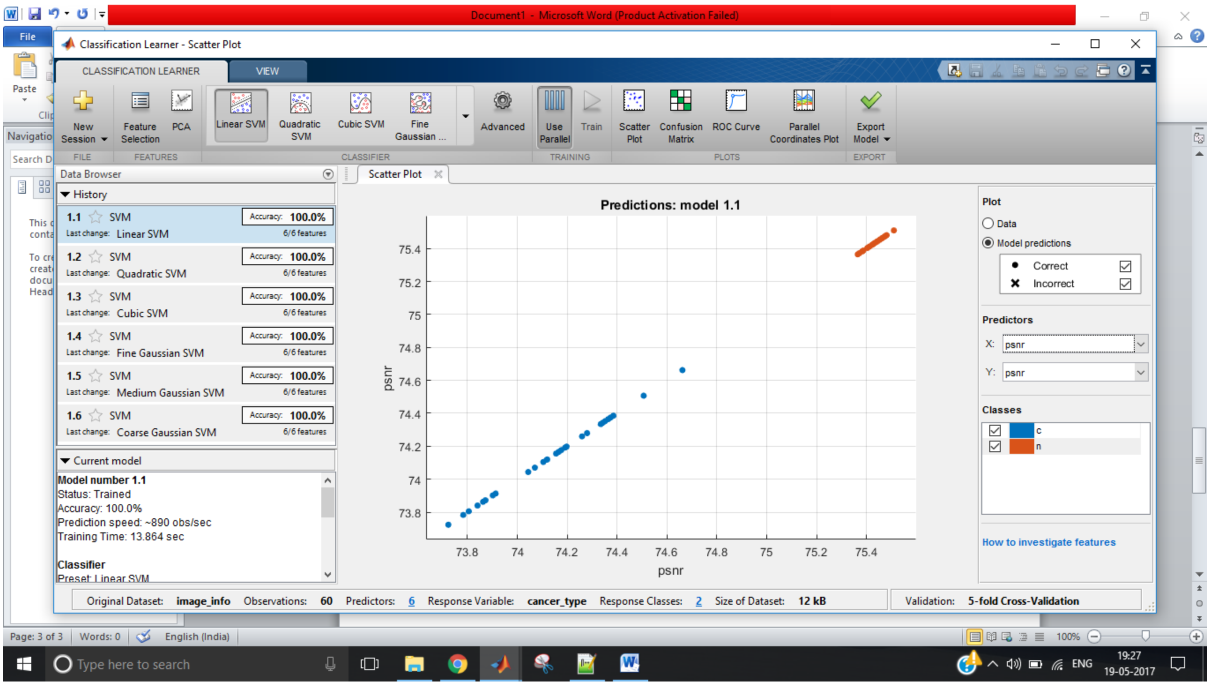
GLCM and KNN

Button Group
☑ Contrast 0.6723
☑ Corelation 1.1021
☐ Energy 0
☑ Homogenity 0.9760
☐ Entropy 0
☐ PSNR 0

Classification Result
Predicted Groups **malignant**

Fig 4.7 KNN Classification Output

Simple SVM GUI



Document1 - Microsoft Word (Product Activation Failed)

Classification Learner - Scatter Plot

Data Browser

History

1.1 SVM Accuracy: 100.0% Last change: Linear SVM 6/6 features
1.2 SVM Accuracy: 100.0% Last change: Quadratic SVM 6/6 features
1.3 SVM Accuracy: 100.0% Last change: Cubic SVM 6/6 features
1.4 SVM Accuracy: 100.0% Last change: Fine Gaussian SVM 6/6 features
1.5 SVM Accuracy: 100.0% Last change: Medium Gaussian SVM 6/6 features
1.6 SVM Accuracy: 100.0% Last change: Coarse Gaussian SVM 6/6 features

Current model

Model number 1.1
Status: Trained
Accuracy: 100.0%
Prediction speed: ~890 obs/sec
Training Time: 13.864 sec

Classifier
Preset: Linear SVM

Original Dataset: image_info Observations: 60 Predictors: 6 Response Variable: cancer_type Response Classes: 2 Size of Dataset: 12 kB Validation: 5-fold Cross-Validation

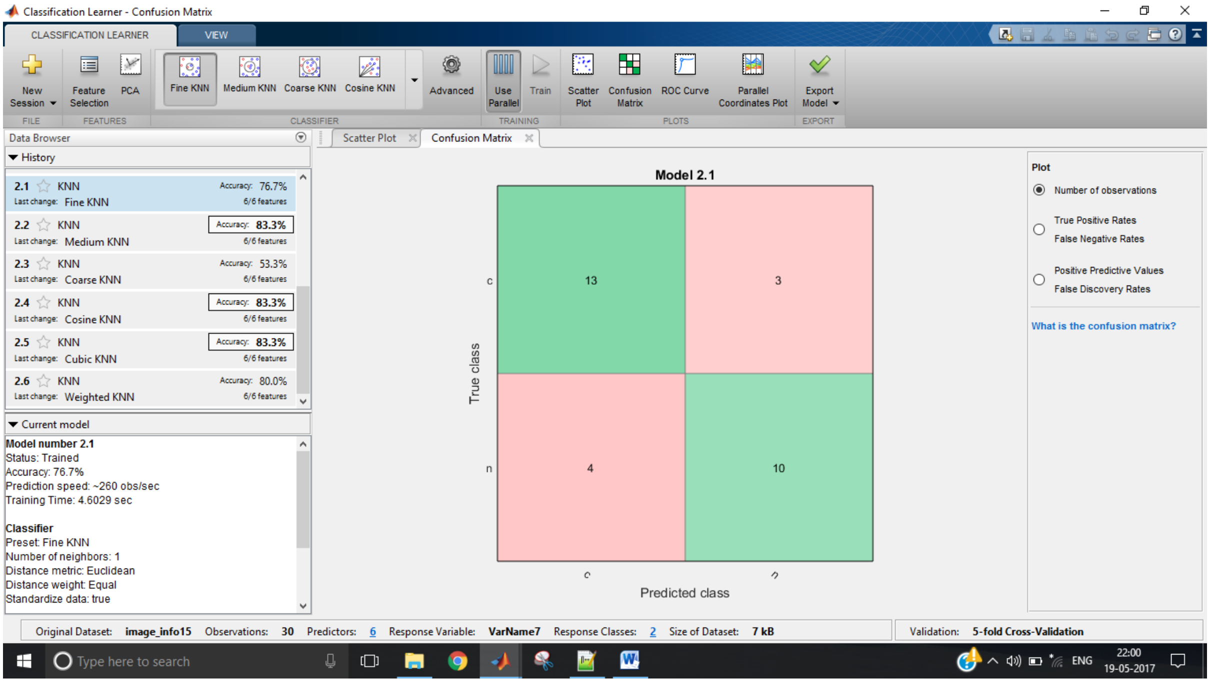
Plot
○ Data
● Model predictions
● Correct ☑
× Incorrect ☑

Predictors
X: psnr
Y: psnr

Classes
☑ c
☑ n

Fig 4.8 SVM Output

Confusion Matrix



2.1 KNN Accuracy: 76.7%
Last change: Fine KNN 6/6 features

2.2 KNN Accuracy: 83.3%
Last change: Medium KNN 6/6 features

2.3 KNN Accuracy: 53.3%
Last change: Coarse KNN 6/6 features

2.4 KNN Accuracy: 83.3%
Last change: Cosine KNN 6/6 features

2.5 KNN Accuracy: 83.3%
Last change: Cubic KNN 6/6 features

2.6 KNN Accuracy: 80.0%
Last change: Weighted KNN 6/6 features

Model number 2.1
Status: Trained
Accuracy: 76.7%
Prediction speed: ~260 obs/sec
Training Time: 4.6029 sec

Classifier
Preset: Fine KNN
Number of neighbors: 1
Distance metric: Euclidean
Distance weight: Equal
Standardize data: true

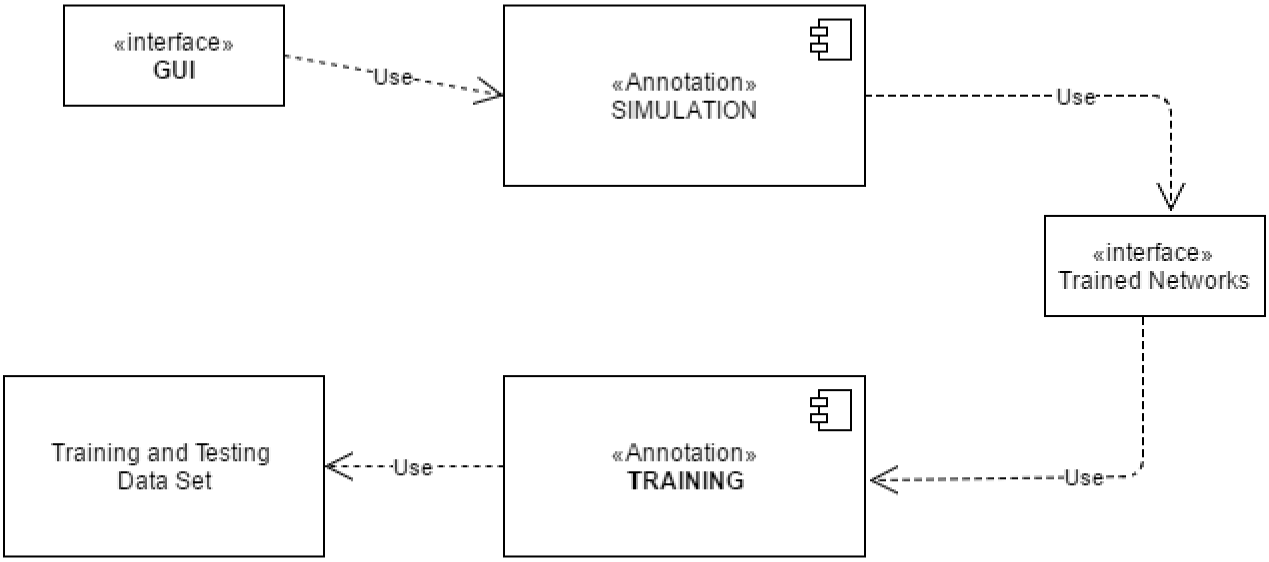
● Number of observations
○ True Positive Rates
False Negative Rates
○ Positive Predictive Values
False Discovery Rates

What is the confusion matrix?

Original Dataset: image_info15 Observations: 30 Predictors: 6 Response Variable: VarName7 Response Classes: 2 Size of Dataset: 7 kB Validation: 5-fold Cross-Validation

Fig 4.10 Confusion Matrix Output

4.4.4 Component Diagram:



Chapter 5. Project Implementation

5.1 Approach/System Architecture / Main Algorithm / Methodology

We have implemented four algorithms. The detailed explanation and outputs are shown below:

I] SEGMENTATION:

1. Fuzzy C-Means Clustering Algorithm

The fuzzy logic is used to process data by giving the partial membership value to each pixel in the image. The membership value of the fuzzy set ranges between 0 and 1. Fuzzy clustering is a multi-valued logic system that uses intermediate values i.e., member of one fuzzy set can also be member of another fuzzy set while in the same image. There are no discontinuous or sudden transitions between full membership and non-membership functions. The membership function defines the fuzziness of an image and the information contained in the image. The primary features involved in characterization using a membership function are: core, support, and boundary. The core is completely a member of the fuzzy set. The support is non membership value of the set and boundary is the partial membership value, having its value between 0 and 1. This algorithm works by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and the data point. More the data is near to the cluster centre, more is its membership towards the particular cluster centre. Hence, addition of membership of each and every data point must be equal to one.

Mathematical representation

Algorithmic steps for Fuzzy c-means clustering

A. Fuzzy Clustering

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centres.

1) Randomly select ‘c’ cluster centres.

2) calculate the fuzzy membership 'µij' using:

3) compute the fuzzy centres 'vj' using:

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or ||U(k+1) - U(k)|| < β.

Where,

‘k’ is the iteration step.

‘β’ is the termination criterion between [0, 1].

‘U = (µij)n*c’ is the fuzzy membership matrix.

‘J’ is the objective function.

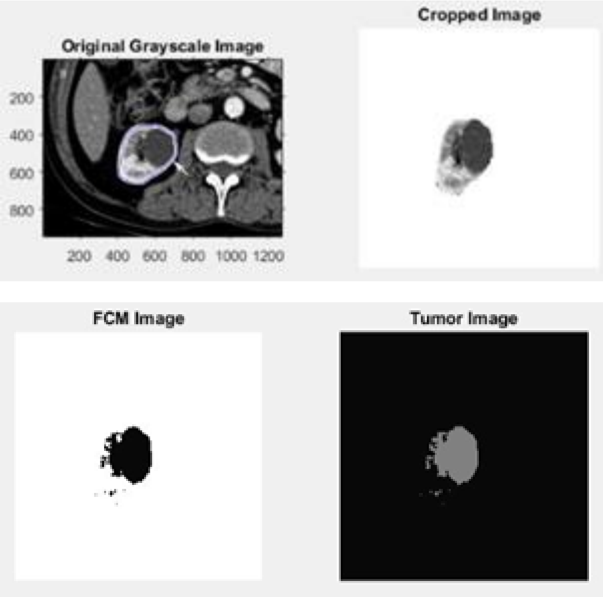


Fig 5.1 Output image for segmentation using FCM.

II] FEATURE EXTRACTION:

2. Gray-Level Co-occurrence Matrix Algorithm

GLCM is a widely used method for medical image analysis, classification. This method gives us information about relative position of two pixels with respect to each other. The GLCM is then created by counting the number of occurrences of pixel pairs at a certain distance. To compute the GLCM matrix for an image f (i, j), a distance vector d=(x, y) is defined. The (i,j)th element of the GLCM matrix P is defined as the probability that grey levels i and j occur at distance d and angle θ, then extracting texture features from GLCM matrix P. Four angles(0,45,90,135)and four distances(1,2,3,4) can be used to calculate the co-occurrence matrix as shown in Fig 3.

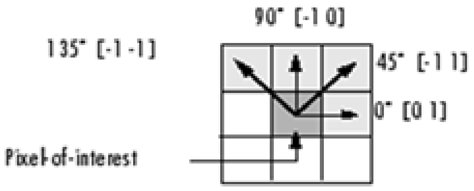


Fig 3:Calculation of co-occurrence matrix in GLCM.

Expressions of GLCM descriptors are:

1) **Correlation** defined by Eq. (1)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i,j)X(i\ X\ j) - (\mu_x X \mu_y)/\sigma_x \sigma_y \qquad ...(1)$$

2) **Contrast** defined by Eq. (2)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i,j)(i-j)^2 \qquad ...(2)$$

3) **Energy** defined by Eq. (3)

$$=\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}P(i,j)^2 \quad ...(3)$$

4) **Entropy** defined by Eq. (4)

$$= - \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}P(i,j)\log(P(i,j)) \quad ...(4)$$

5) **Homogeneity** defined by Eq. (5)

$$=\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{p(i,j)}{1+|i-j|} \quad ...(5)$$

6) **Peak Sound to Noise Ratio** (PSNR)

The PSNR value calculates the peak signal-to-noise ratio, in decibels, between two images. This ratio is then used as a quality measurement between the original image and a compressed image. The higher the PSNR value, better is the quality of the compressed or reconstructed image. The Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are the two error metrics that are used to differentiate amongst two image compression qualities. The MSE value represents the cumulative squared error between the compressed and the original image, and the PSNR value represents measure of the peak error. The lower the value of MSE, lower is the error. To compute the PSNR, the block first calculates the mean-squared error using the following Eq.(6):

$MSE=M,N[I1(m,n)-I2(m,n)]^\wedge 2/M*N$ …(6)

In the previous equation, M and N are the number of rows and columns in the input images, respectively. Then the block computes the PSNR using the following Eq. (7):

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \quad ..(7)$$

III] CLASSIFICATION

3. K-Nearest Neighbour

The k-nearest neighbor is a semi-supervised learning algorithm. It requires training data and a predefined k value to find the k nearest data based on distance computation. If k data have different classes, the algorithm predicts class of the unknown data to be the same as the majority class.

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x1, x2, ..., xmx, and my-by-n data matrix Y, which is treated as my(1-by-n) row vectors y1, y2,..., ymy, the various distances between the vectors xs and yt are defined as follows:

1. Euclidean Distance

The Euclidean distance is a measure to find

distance between two points, defined by Eq. (8)

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)'$$

...(8)

2.Hamming Distance

Hamming distance, which is the percentage of coordinates that differ [8], can be defined by Eq.(9)

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right)$$

...(9)

3.Cosine Distance

The Cosine distance is computed from one minus the cosine of the included angle between points [8], defined by Eq. (10)

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right)$$

...(10)

4. City Block Distance[8]

The city block distance between two points is the summation of the absolute difference of Cartesian coordinates, defined by Eq. (11)

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}|$$

...(11)

5. Correlation Distance

Distance based on correlation is a measure of

statistical dependence between two vectors, defined by

Eq. (12)

$$d_{st} = \left(1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}}\right)$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}$$

...(12)

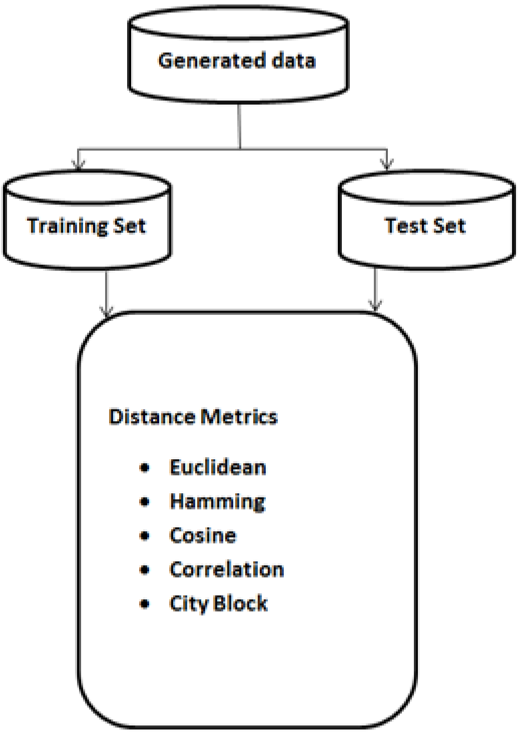
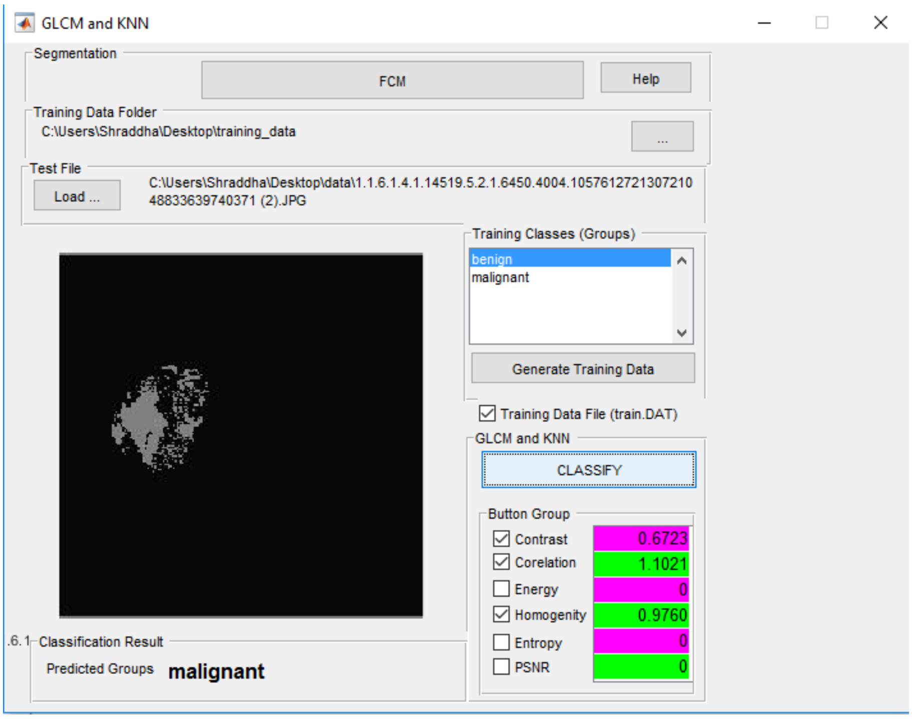


Fig 5.2 Classification data flow

Step 1: Generate a binary data set with several different distribution and different amount of data in each class.

Step 2: Use data from step 1 for data classification by applying the k-nearest neighbour algorithm with various distance metrics to compute the k-nearest data points for making classification.

Step 3: Analyse the results and conclude about the performance of classification using various distance metrics, as seen for classifying tumor as malignant or benign in Fig 5.



GLCM and KNN

Training Data Folder
C:\Users\Shraddha\Desktop\training_data

Test File
C:\Users\Shraddha\Desktop\data\1.1.6.1.4.1.14519.5.2.1.6450.4004.105761272130721048833639740371 (2).JPG

Training Classes (Groups)
benign
malignant

☑ Training Data File (train.DAT)

☑ Contrast	0.6723
☑ Corelation	1.1021
□ Energy	0
☑ Homogenity	0.9760
□ Entropy	0
□ PSNR	0

.6.1 Classification Result
Predicted Groups **malignant**

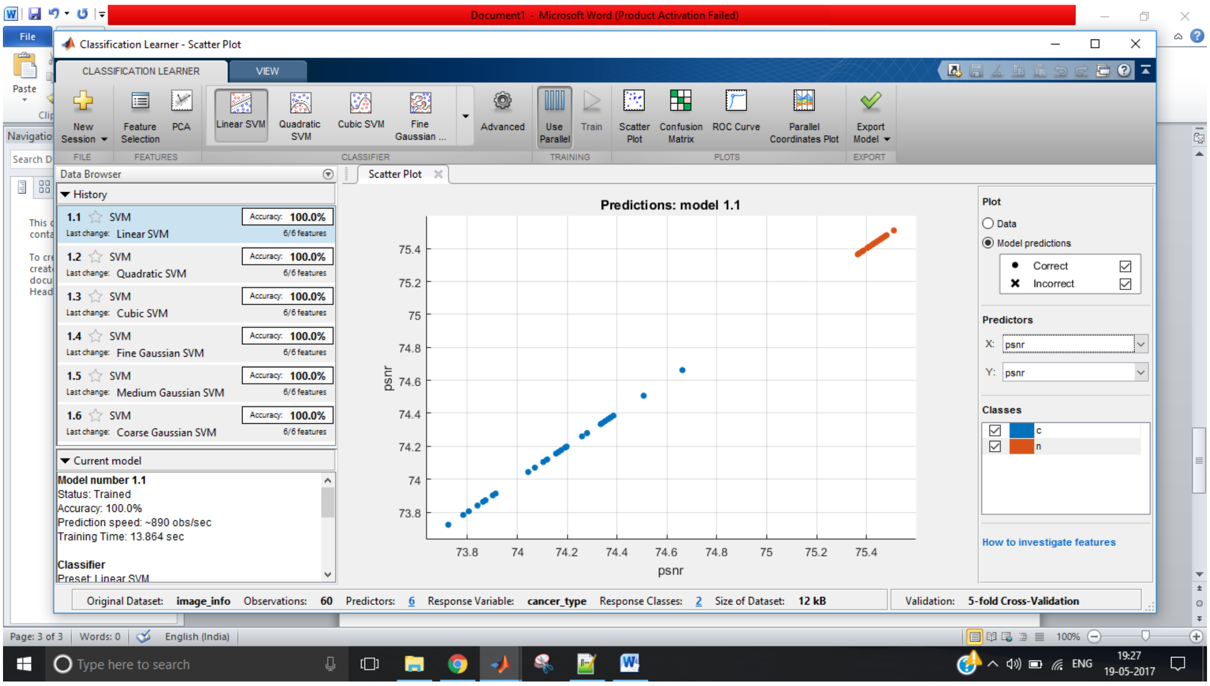
Fig 5.3 KNN output

Classification using KNN

4. Support Vector Machine

Support vector machines (SVMs) are a type of supervised learning models along with associated learning algorithms that analyze data and recognize various patterns, used for classification analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes, malignant and benign forms the output, making it a non-probabilistic binary linear classifier. Now that there are set of training examples at hand, each marked as belonging to one of two categories, an SVM training algorithm constructs a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Newer examples are then plotted into it and then predicted to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. A clear segregation of the two types of data points in Fig 6 is seen, showing the hyperplane generated using SVM.



1.1	SVM	Accuracy: 100.0%	
	Last change: Linear SVM		6/6 features
1.2	SVM	Accuracy: 100.0%	
	Last change: Quadratic SVM		6/6 features
1.3	SVM	Accuracy: 100.0%	
	Last change: Cubic SVM		6/6 features
1.4	SVM	Accuracy: 100.0%	
	Last change: Fine Gaussian SVM		6/6 features
1.5	SVM	Accuracy: 100.0%	
	Last change: Medium Gaussian SVM		6/6 features
1.6	SVM	Accuracy: 100.0%	
	Last change: Coarse Gaussian SVM		6/6 features

Model number 1.1
Status: Trained
Accuracy: 100.0%
Prediction speed: ~890 obs/sec
Training Time: 13.864 sec

Classifier
Preset: Linear SVM

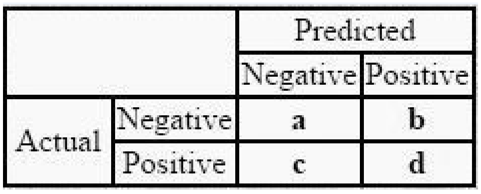
Original Dataset: **image_info** Observations: **60** Predictors: 6 Response Variable: **cancer_type** Response Classes: 2 Size of Dataset: **12 kB** Validation: **5-fold Cross-Validation**

Fig 5.4 Scatter Plot of SVM.

Confusion Matrix

The output of the computing done via SVM is mapped via the confusion matrix. A confusion matrix consist of information about actual and predicted classes done by a classification system. Performance of such systems is evaluated using the data in the matrix. The following table in Fig 7 shows the confusion matrix for a two class classifier.Confusion Matrix helps in detecting the accuracy of datasets.

The data in the confusion matrix as shown in Fig 7 have the following meaning in context.



		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Table 5.1 Confusion Matrix Calculation

a is the number of correct predictions that an instance is negative,

b is the number of incorrect predictions that an instance is positive,

c is the number of incorrect of predictions that an instance negative, and

d is the number of correct predictions that an instance is positive.

5.2 Programming Language used for Implementation

MATLAB

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. A proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages.

5.3 Tools used

1.Neural Network ToolBox.
2. Languages: MATLAB.
3. Documentation: Microsoft Word, Google docx.
4. Software: MATLAB 2016a.
5. Hardware: Intel Dual Core Dual Processor or advanced version; Minimum 256 MB of RAM; Minimum 1 GB of Hard disk Space

5.4 Deployment Diagram:

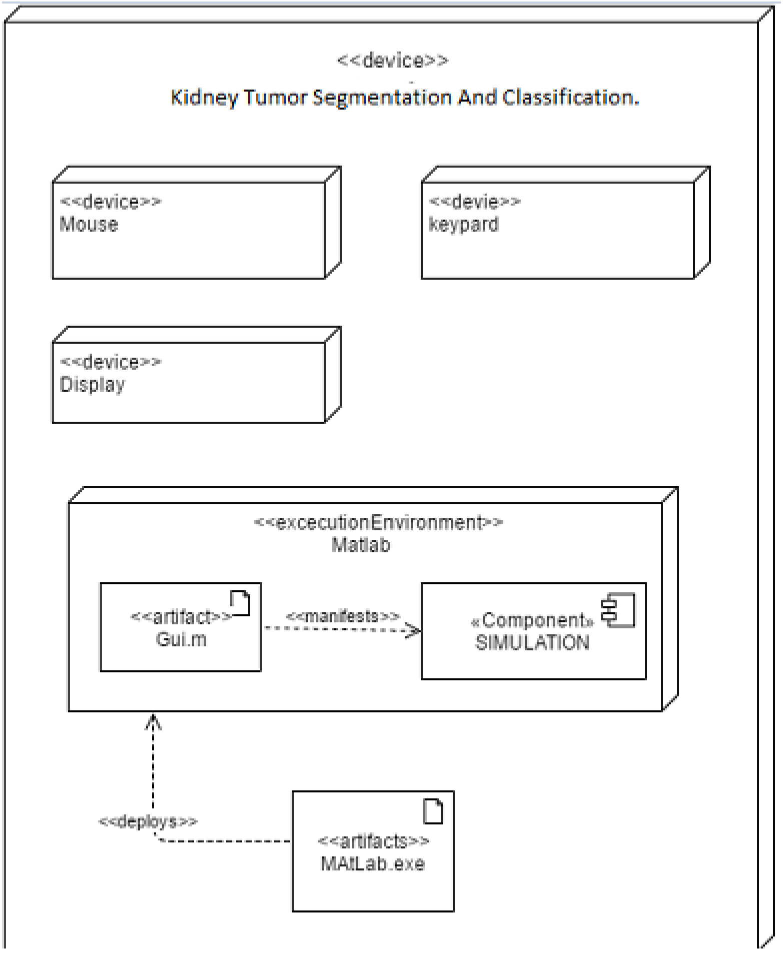


Fig 5.5 Deployment Diagram

Chapter 6. Integration & Testing

6.1 Testing Approach

· **Unit testing**

- It concentrates on the efforts required for verification on the minute units of software design which namely, is the software module.
- We use the component level design description as a guide. Important control paths are tested to uncover errors within the boundary of the module.
- The unit test is white box oriented, and the steps can be conducted in parallel for multiple modules.

· **Integration Testing**

- Interfacing of various modules can be problematic.
- Data loss can occur across an interface, one module may affect the other, and individually acceptable imprecision may be magnified when combined.
- Integration testing is thus used to construct the program structure and also to conduct tests to unfold errors related to the interface.

· **Stress testing**

- Stress tests are conducted to confront programs with abnormal situations.
- Stress testing forces a system to execute in a manner that demands for resources in abnormal quantity, frequency or volume which allows us to gauge the limit of the system.

· **Performance Testing**

- Performance testing is conducted through all the steps in the testing process to test runtime performance of software within the context of an integrated system.

· **Security Testing**

- This system manages sensitive information related to patients. There may be causes and actions that can harm these individuals thus becoming a target for improper or illegal penetration.
- Security testing attempts to verify that protection mechanisms built into a system will, in fact, protect it from improper penetration.
- During security testing, the tester plays the role of the hacker who desires to penetrate the system.
- Given enough time and resources, good security testing will ultimately penetrate the system. The role of the system designer is make penetrate cost more than value of the information that will be obtained.

· **Recovery Testing**

- Many computers – based systems must recover from faults and resume processing within a pre – specified time. In some cases, a system must be fault tolerant, i.e. processing faults must not cause overall system function to cease. In other cases, a system failure must be corrected within a specified period of time or severe economic damage will occur.
- The testing approach that was followed in the project began at the preliminary level and this testing was worked outwards toward the integration of the entire system.
- Testing approach was an umbrella activity in the development of the system. Each module was tested at its completion before transitioning to the next component.
- After selecting a particular data set for a kidney tumor and selecting some probably suitable combinations for the network, they were tested and the output was compared with actual desired output.
- After the development of GUI the integration of network was tested with the GUI.
- Different types of testing adopted in the approach are as follows: unit testing, Integration testing.

6.2 Testing Plan

Unit testing:

The significance of Unit Testing in this project was prominent since every algorithm had its own unique configuration and best case scenario for the same data set was to be found. For different algorithm , the algorithm were trained and tested and the result was obtained as a confusion matrix to check the accuracy of sample data set .

Integrated testing:

Integration testing is used to check integration of the final selected algorithm with the GUI. Tests were done after integrating the said GUI with the algorithm.

Testing schedule:

Test Title	**Date**
Test the selected data set with Fuzzy C-Means algorithm.	10th August 2016
Test the selected data set with KNN algorithm.	30th October 2016
Test the selected data set with SVM algorithm.	30th January 2016
Extensive testing of the best selected classification algorithms	1st February 2017
Integrated testing with the GUI	2nd March 2017

Table 6.1: Testing Schedule

6.3 Unit Test Cases:

Test case:	**1**	**2**	**3**	**4**
Total number of data records:	10	15	36	51
KNN Accuracy	45%	53%	62%	83%
SVM Accuracy	50%	76%	95%	100%

Table 6.2 Unit Test Cases

6.4 Integrated System Test Cases:

The testing results remained same as the unit test cases even after integrating it with the GUI, as the backend process remained the same only the front end showed the latency depicting the training being done behind.

Chapter 7: Conclusion & Future work

Various segmentation and classification of kidney and kidney tumor methods for CT image have been discussed in this paper. This performance study reveals that semi automated segmentation method reduces errors occurring while doing manual segmentation. Experimental results were obtained by using matlab software. In this paper improvement in the accuracy of results are obtained by incorporating noise removal steps while doing fuzzy c means clustering. Final stage in tumor detection is to decide whether it is cancerous or non-cancerous. This classification is done by designing an artificial neural network using KNN, and also implemented using SVM algorithm. The accuracy is mapped using the confusion matrix, and the outcomes are compared. Improvement in algorithm is done by training the algorithm with more image features and by doing rigorous testing.

References

[1] Anton Bardera, Jaume Rigau, Imma Boada, Miquel Feixas, and Mateu Sbert, " Image Segmentation Using Information Bottleneck Method ",Page Number 1601-1612, IEEE Transactions on Image Processing, Vol. 18, No. 7, July 2009.

[2] Mr.Rohit S. Kabade and Dr. M. S. Gaikwad, "Segmentation of Brain Tumour and Its Area Calculation in Brain MR Images using K-Mean Clustering and Fuzzy CMean Algorithm " IJCSET Vol. 4 No. 05 May 2013

[3] J.selvakumar ,A.Lakshmi and T.Arivoli."Brain Tumor Segmentation and Its Area Calculation in Brain MR Images using K-Mean Clustering and Fuzzy C-Mean Algorithm"IEEE- ICAESM -2012 March 30, 31, 2012

[4]Robert.L.Cannon;Jitendra.V.Dave;James.C.Bezdek"Efficient Implementation of the Fuzzy c-Means Clustering Algorithms"IEEE TRANSACTFIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. VOL. PAMI-8, NO. 2, MARCH 1986

[5]Kuldeep Pawar , Jayamala.K.Patil "A Systematic Study of Segmentation Methods For Detection of Kidney Tumor Using Computed Tomography Images."International Journal of Engineering Sciences & Research Technology Nov., 2012

[6] Mredhula.L and Dr.M.A.Dorairangaswamy "Detection And Classification of Tumors in CT Images" IJCSE Vol. 6 No.2 Apr-May 2015

[7]U Akilandeswari, Nithya Rajendran and B Santhi, "Review on Feature Extraction Methods in Pattern Classification", European Journal of Scientific Research, Vol.71 No.2 (2012).

[8] Kittipong Chomboon* , Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, Nittaya Kerdprasop, “An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm “, International Conference on Industrial Application Engineering 2015

[9]Jian Wu, Feng Ye, Jian-Lin Ma, Xiao-Ping Sun, Jing Xu, Zhi-Ming, “ The Segmentation and Visualization of Human Organs Based on Adaptive Region Growing Method” ,Page Number-439-443, IEEE 8th International Conference on Computer and Information Technology Work shops978-0-7695-3242-4/08,IEEE,2008.

[10] Rafael C. Gonzalez, Richard E.Woods, “ Digital Image processing”, published by Pearson Education, Inc. 2002.

[11] Kuldeep Pawar, Jayamala K Patil, Bharathi Vidyapeeth’s College of Engineering, Kolhapur, India, “A Systematic Study of Segmentation Methods For Detection of Kidney Tumor Using Computed Tomography Images” IJESRT [Pawar, 1(10): Nov., 2012]

[12] Bansari Shah, Charmi Sawla, Shraddha Bhanushali, Poonam Bhogale Kidney Tumor Segmentation and Classification on Abdominal CT Scans. *International Journal of Computer Applications* 164(9):1-5, April 2017

Appendix

I) Minimum System Requirement

Software Requirements:

- MATLAB 2016a
- ONLINE DATABASE:

www.cancerimagingarchive.net and database collected from pathology and radiology labs

- Microsoft Office

Hardware Requirements:

- Intel Dual Core Dual Processor or advanced version
- Minimum 8GB of RAM
- Minimum 1 GB of Hard disk Space

II) User’s Manual

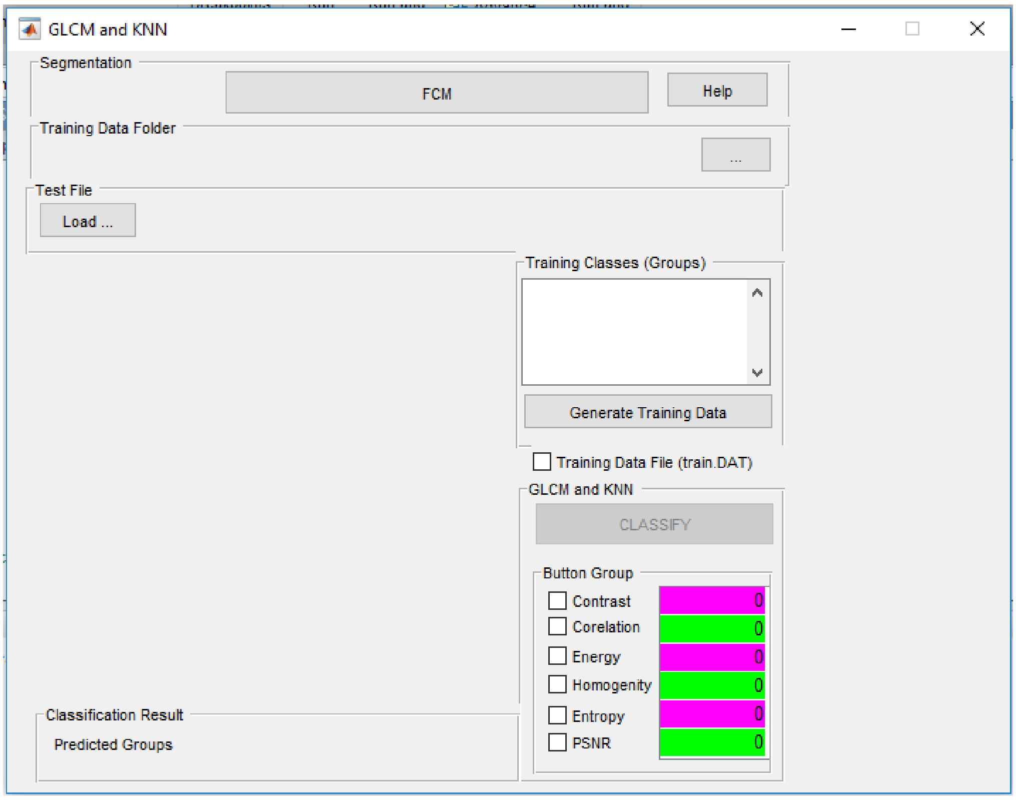
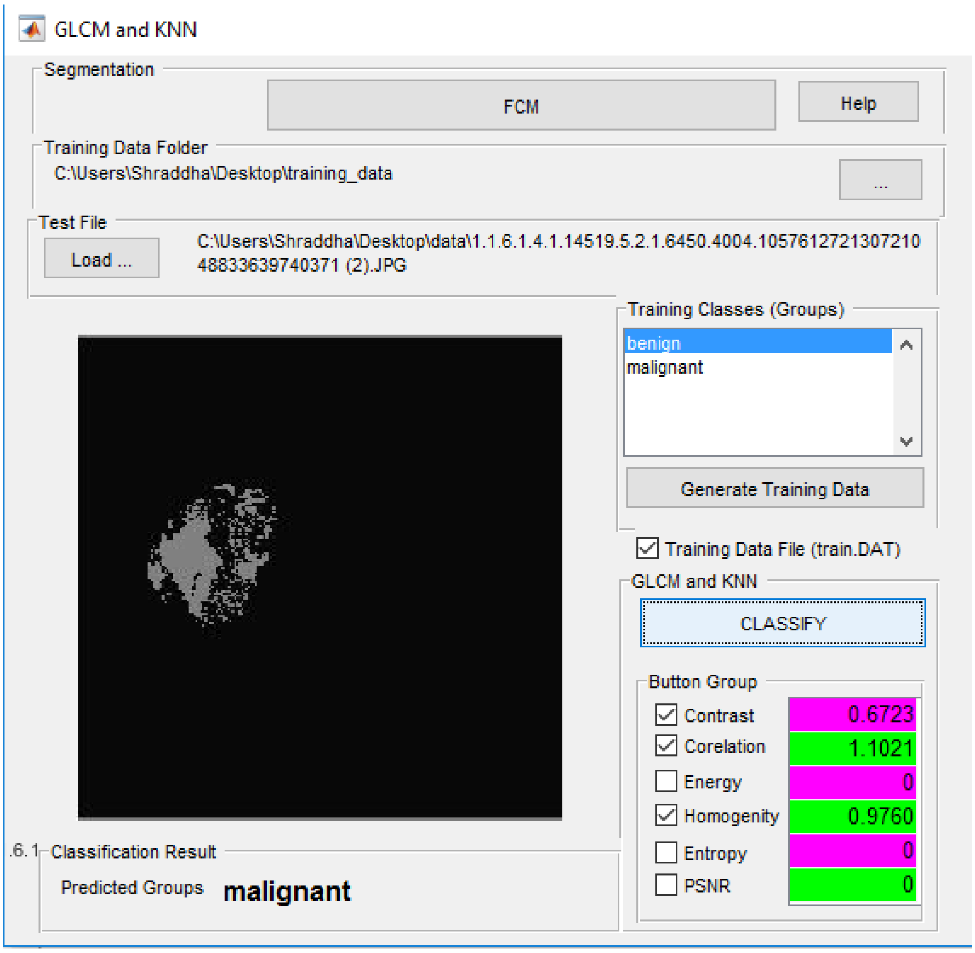


Fig 7.1 GUI for user

1. Load folder containing tumor image training data folders with classified data to used for training algorithm using ‘…’ button beside ‘Help’ by browsing through device files
2. Check if respective training classes have been displayed under the ‘Training Classes (Groups)’ list
3. Application will generate file ‘groups.dat’ in training folder and the info of sub folders will be inside this file
4. Generate training data by clicking on ‘Generate Training Data’ button below the list
5. User will select the features it wants to compare during feature extraction using the checkboxs at the lower right of the UI
6. Application will generate file ‘train.dat’ in training folder and the GLCM features of given classes will be in this file
7. Select a JPG/JPEG file by browsing through system files, pertaining to training data to test using ‘Load...’ button
8. Check if correct image is loaded and displayed at centre screen space
9. Click ‘Classify’ button
10. GLCM features of image will be displayed underClassify button in a tabular format

11. This application (during classification process) will generate three result files:
    a. Copy of Original JPEG/JPG file what we use as test data
    b. data file as features analysed of target image
    c. KNN result = sorted measured distance for each K in the number of K (depending on the number of class)
12. The nearest distance in result is given category (displayed in knn_dat file
13. Predicted group is displayed at the bottom



Training Data Folder
C:\Users\Shraddha\Desktop\training_data

Test File
C:\Users\Shraddha\Desktop\data\1.1.6.1.4.1.14519.5.2.1.6450.4004.1057612721307210 48833639740371 (2).JPG

Training Classes (Groups)
benign
malignant

☑ Training Data File (train.DAT)

GLCM and KNN

Button Group

☑ Contrast	0.6723
☑ Corelation	1.1021
□ Energy	0
☑ Homogenity	0.9760
□ Entropy	0
□ PSNR	0

Classification Result
Predicted Groups **malignant**

Fig 7.2 GUI of output for user

The system is user-friendly. There is hardly any requirement of User Manual. Users just have to proceed with the software by clicking the buttons and the system will fulfill the user’s requirements.

III) Technical Reference Manual

Installing MATLAB is requirement for implementation of this system. Licensed version of the software is obtained from the internet and next we have to follow the instructions given in their user manual. Images loaded must be of desired format (JPEG/JPG), and must not exceed 640*640 dimension

Papers Published

Bansari Shah, Charmi Sawla, Shraddha Bhanushali and Poonam Bhogle. Kidney Tumor Segmentation and Classification on Abdominal CT Scans. *International Journal of Computer Applications* 164(9):1-5, April 2017.

Paper Presented

‘Kidney Tumor Segmentation and Classification on Abdominal CT Scans’ for a state level conference and working model exhibition competition, Prakalp’17 presented by Indian Society for Technical Education.on 31 st March, 2017.

Copy of published paper

International Journal of Computer Applications (0975 – 8887)
Volume 164 – No 9, April 2017

Kidney Tumor Segmentation and Classification on Abdominal CT Scans

Bansari Shah
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Charmi Sawla
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Shraddha Bhanushali
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Poonam Bhogale
Assistant Professor of Computer Engineering
K J Somaiya COE
Vidyavihar

ABSTRACT

This paper, deals with systematic study of simple segmentation and classification algorithms for kidney tumor using Computed Tomography images. Tumors are of different types having different characteristics and also have different treatment. It becomes very important to detect the tumor and classify it at the early stage so that appropriate treatment can be planned. This CT scans are visually examined by the physician for detection and diagnosis of kidney tumor. However this method lacks accuracy and detection of size of the tumor. So to overcome this, a computer aided segmentation technique has been proposed, which extracts the tumor part from the kidney, further on which feature extraction method is performed for extracting certain features and the type of tumor i.e. malignant or benign is displayed by using simple classifiers .

General Terms

Algorithms, Kidney Tumor, Computed Tomography scans, Process.

Keywords

Pre-processing, Fuzzy C-means, Grey Level Co-occurrence Matrix, K Nearest Neighbour classifier, Support Vector Machine classifier

1. INTRODUCTION

Human body consists of myriad number of cells[5]. For a body to remain healthy, cells grow and divide in orderly fashion. When cell growth becomes uncontrollable the extra mass of cell transforms into tumor. CT scans and MRI are used for identification of tumor[5]. The goal of the system is to detect tumor by incorporating image processing, pattern analysis, and computer vision techniques for enhancement, segmentation and classification for kidney diagnosis, as shown in Fig 1.. This system can be used by radiologists and healthcare specialists[1,2,9,10]. The system is expected to improve the sensitivity, specificity, and efficiency of kidney tumor screening. Tumor segmentation is done by fuzzy c-means algorithm , extraction of features is implemented by grey level co-occurrence matrix method (GLCM) and finally classification of tumor if it is benign or malignant is obtained by support vector machine tool (SVM) and k nearest neighbour classifier (KNN).

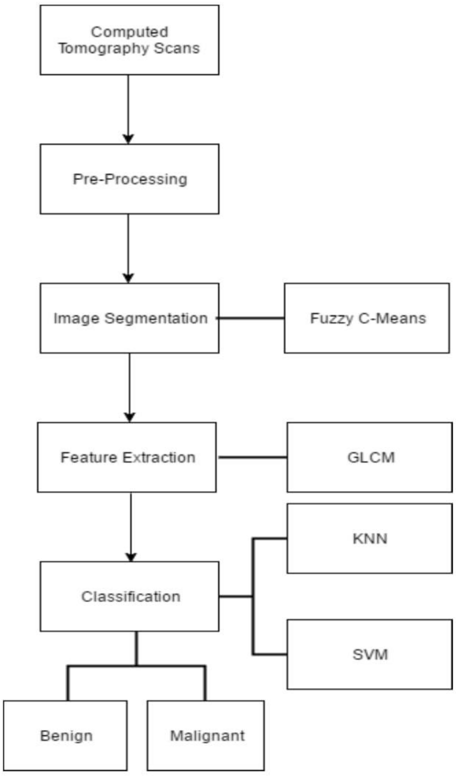


Fig 1: Block Diagram Of Proposed System

2. SEGMENTATION

2.1 Fuzzy C-Means Clustering Algorithm

A. Fuzzy Clustering

The fuzzy logic is used to process data by giving the partial membership value to each pixel in the image. The membership value of the fuzzy set ranges between 0 and 1. Fuzzy clustering is a multi valued logic system that uses intermediate values i.e., member of one fuzzy set can also be member of another fuzzy set while in the same image. There are no discontinuous or sudden transitions between full membership and non-membership functions. The membership function defines the fuzziness of an image and the information contained in the image. The primary features involved in characterization using a membership function are: core, support, and boundary. The core is completely a member of the fuzzy set. The support is non membership value of the set and boundary is the partial membership value, having its value between 0 and 1. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center, more is its membership towards the particular cluster center. Hence, addition of membership of each and every data point must be equal to one.

B. Mathematical representation

Algorithmic steps for Fuzzy c-means clustering [2]

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the fuzzy membership 'µij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3) Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or ||U(k+1) - U(k)|| < β.
where,

'k' is the iteration step.
'β' is the termination criterion between [0, 1].
'U = (µij)n*c' is the fuzzy membership matrix.
'J' is the objective function.

C. Screenshot for Fuzzy C-Means Clustering

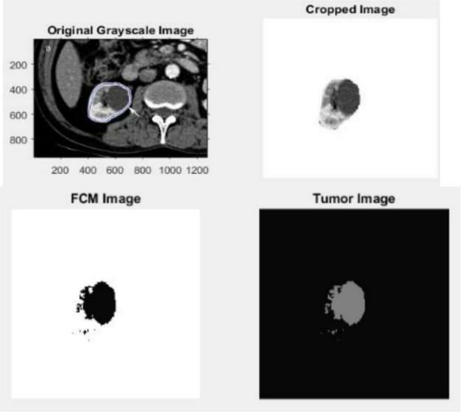


Fig 2: Output image for segmentation using FCM.

Fig.2 is the image of tumor obtained after cropping kidney from abdominal scan and applying Fuzzy C-Means Clustering to it. The Fuzzy C-Mean algorithm clusters the image according to some characteristics.

3. FEATURE EXTRACTION

3.1 Grey Level Co-occurrence Matrix

A.GLCM

GLCM is a widely used method for medical image analysis, classification. This method gives us information about relative position of two pixels with respect to each other. The GLCM is then created by counting the number of occurrences of pixel pairs at a certain distance. To compute the GLCM matrix for an image f (i, j), a distance vector d=(x, y) is defined. The (i,j)th element of the GLCM matrix P is defined as the probability that grey levels i and j occur at distance d and angle θ, then extracting texture features from GLCM matrix P. Four angles(0,45,90,135)and four distances(1,2,3,4) can be used to calculate the co-occurrence matrix as shown in Fig 3.

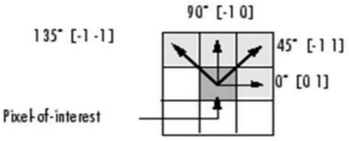


Fig 3:Calculation of co-occurrence matrix in GLCM.

B. Expressions of GLCM descriptors are: [7]

1) Correlation defined by Eq. (1)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i, j) X (i \; X \; j) - (\mu_x X \mu_y) / \sigma_x \sigma_y \quad ...(1)$$

2) Contrast defined by Eq. (2)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i, j)(i - j)^2 \quad ...(2)$$

3) Energy defined by Eq. (3)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i, j)^2 \quad ...(3)$$

4) Entropy defined by Eq. (4)

$$= - \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} P(i, j) \log(P(i, j)) \quad ...(4)$$

5) Homogeneity defined by Eq. (5)

$$= \Sigma_{i=0}^{G-1} \Sigma_{j=0}^{G-1} \frac{p(i, j)}{1 + | i - j |} \quad ...(5)$$

6) Peak Sound To Noise Ratio(PSNR)

The PSNR value calculatesthe peak signal-to-noise ratio, in decibels, between two images. This ratio is then used as a quality measurement between the original image and a compressed image. The higher the PSNR value, better is the quality of the compressed or reconstructed image.

The Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are the two error metrics that are used to differentiate amongst two image compression qualities. The

MSE value represents the cumulative squared error between the compressed and the original image, and the PSNR value represents measure of the peak error. The lower the value of MSE, lower is the error.

To compute the PSNR, the block first calculates the mean-squared error using the following Eq.(6):

$MSE={}_{M,N}[I_1(m,n)-I_2(m,n)]\hat{}\,_2/M*N$ ...(6)

In the previous equation, M and N are the number of rows and columns in the input images, respectively. Then the block computes the PSNR using the following Eq. (7):

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right)$$

...(7)

4. CLASSIFICATION

4.1 K-Nearest Neighbour

A. KNN

The k-nearest neighbor is a semi-supervised learning algorithm. It requires training data and a predefined k value to find the k nearest data based on distance computation. If k data have different classes, the algorithm predicts class of the unknown data to be the same as the majority class.

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x1, x2, ..., xmx, and my-by-n data matrix Y, which is treated as my(1-by-n) row vectors y1, y2,..., ymy, the various distances between the vectors xs and yt are defined as follows:

1. Euclidean Distance

The Euclidean distance is a measure to find distance between two points [8], defined by Eq. (8)

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)'$$

...(8)

2.Hamming Distance Hamming distance, which is the percentage of coordinates that differ [8], can be defined by Eq.(9)

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right)$$

...(9)

3.Cosine Distance The Cosine distance is computed from one minus the cosine of the included angle between points [8], defined by Eq. (10)

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right)$$

...(10)

4. City Block Distance[8]

The city block distance between two points is the summation of the absolute difference of Cartesian coordinates, defined by Eq. (11)

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}|$$

...(11)

5. Correlation Distance

Distance based on correlation is a measure of statistical dependence between two vectors [8], defined by Eq. (12)

$$d_{st} = \left(1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\,\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}}\right)$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}$$

...(12)

B. Empirical Study Methodology In this section is presented, a study framework using k-nearest neighbor algorithm with various distance metrics. The framework is shown in Fig4.

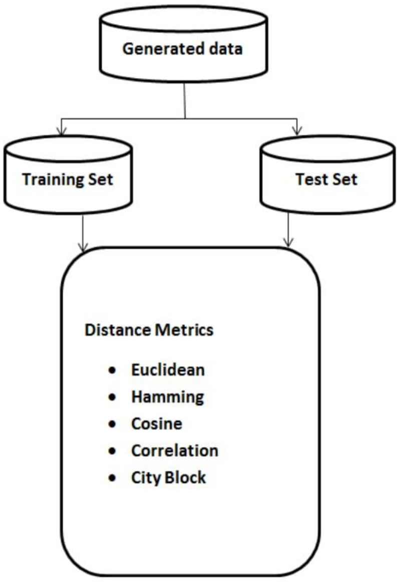


Fig 4:The framework of our empirical study.

From Fig. 4 the detail of each step can be explained as follows:

Step 1: Generate a binary data set with several different distribution and different amount of data in each class. Step 2: Use data from step 1 for data classification by applying the k-nearest neighbor algorithm with various distance metrics to compute the k-nearest data points for making classification. Step 3: Analyze the results and conclude about the performance of classification using various distance metrics, as seen for classifying tumor as malignant or benign in Fig 5.

ANALYSIS AND MACHINE INTELLIGENCE. VOL. PAMI-8, NO. 2, MARCH 1986

[5]Kuldeep Pawar , Jayamala.K.Patil "A Systematic Study of Segmentation Methods For Detection of Kidney Tumor Using Computed Tomography Images."International Journal of Engineering Sciences & Research Technology Nov., 2012

[6] Mredhula.L and Dr.M.A.Dorairangaswamy "Detection And Classification of Tumors in CT Images" IJCSE Vol. 6 No.2 Apr-May 2015

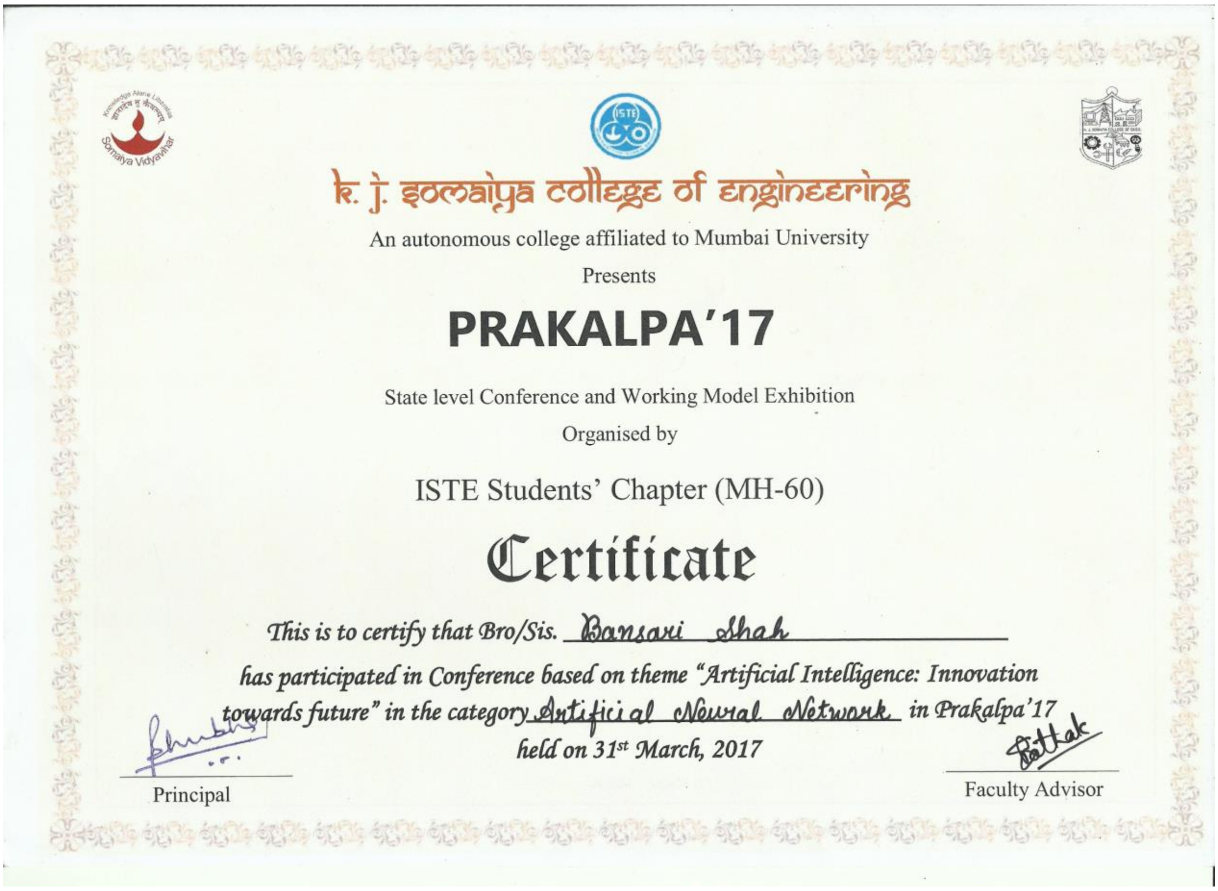
[7]U Akilandeswari, Nithya Rajendran and B Santhi, "Review on Feature Extraction Methods in Pattern Classification", European Journal of Scientific Research, Vol.71 No.2 (2012).

[8] Kittipong Chomboon* , Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, Nittaya Kerdprasop, "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm ", International Conference on Industrial Application Engineering 2015

[9]Jian Wu, Feng Ye, Jian-Lin Ma, Xiao-Ping Sun, Jing Xu, Zhi-Ming, " The Segmentation and Visualization of Human Organs Based on Adaptive Region Growing Method" ,Page Number-439-443, IEEE 8th International Conference on Computer and Information Technology Work shops978-0-7695-3242-4/08,IEEE,2008.

[10] Rafael C. Gonzalez, Richard E.Woods, " Digital Image processing", published by Pearson Education, Inc. 2002.

Participation certificate for project competition



k. j. somaiya college of engineering

An autonomous college affiliated to Mumbai University

Presents

PRAKALPA'17

State level Conference and Working Model Exhibition

Organised by

ISTE Students' Chapter (MH-60)

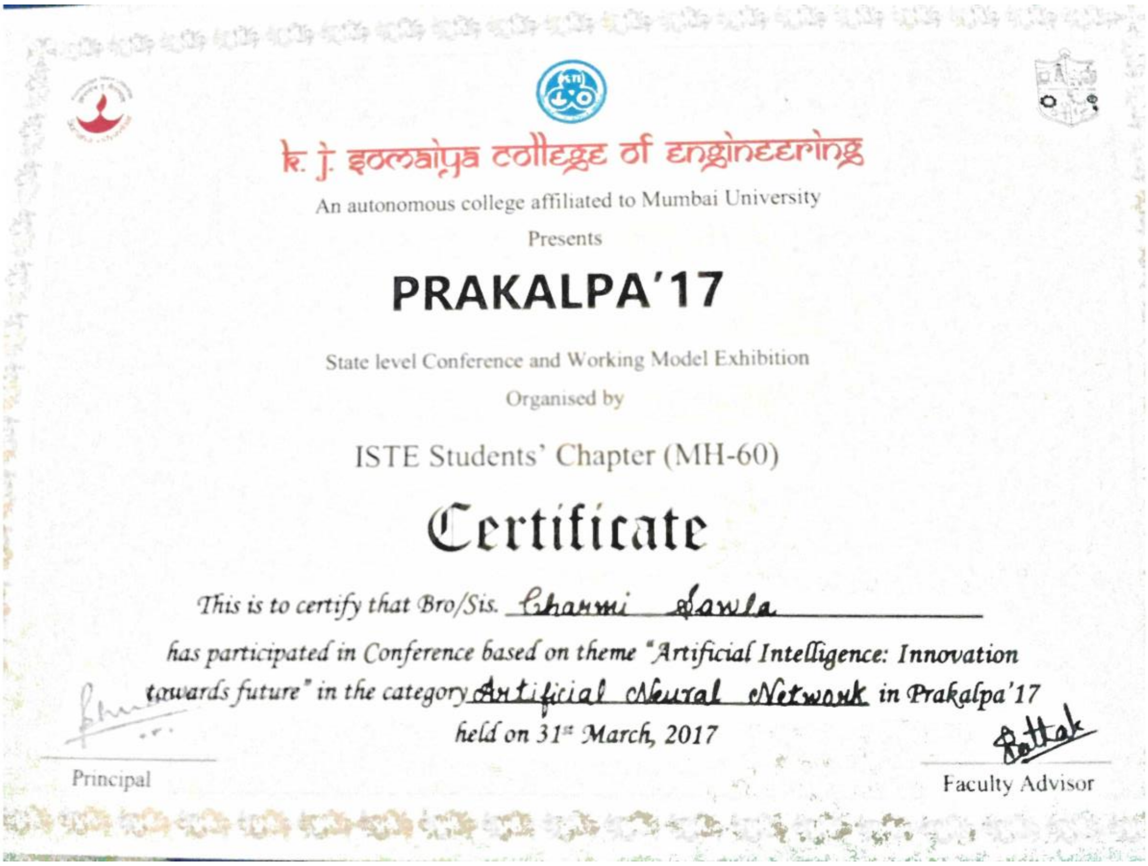
Certificate

This is to certify that Bro/Sis. Bansari Shah

has participated in Conference based on theme "Artificial Intelligence: Innovation towards future" in the category Artificial Neural Network *in Prakalpa'17 held on 31st March, 2017*

Principal

Faculty Advisor



k. j. somaiya college of engineering

An autonomous college affiliated to Mumbai University

Presents

PRAKALPA'17

State level Conference and Working Model Exhibition

Organised by

ISTE Students' Chapter (MH-60)

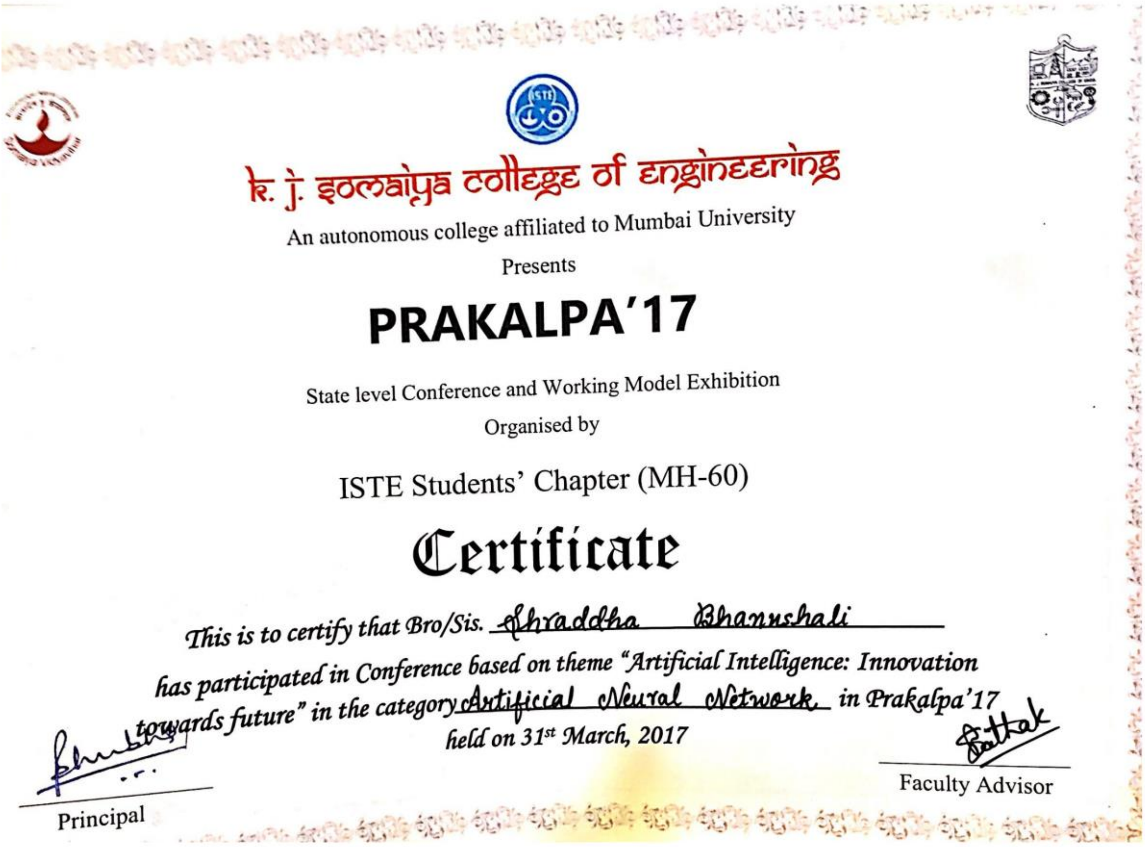
Certificate

This is to certify that Bro/Sis. Charmi Sawla

has participated in Conference based on theme "Artificial Intelligence: Innovation towards future" in the category Artificial Neural Network *in Prakalpa'17 held on 31st March, 2017*

Principal

Faculty Advisor



k. j. somaiya college of engineering

An autonomous college affiliated to Mumbai University

Presents

PRAKALPA'17

State level Conference and Working Model Exhibition

Organised by

ISTE Students' Chapter (MH-60)

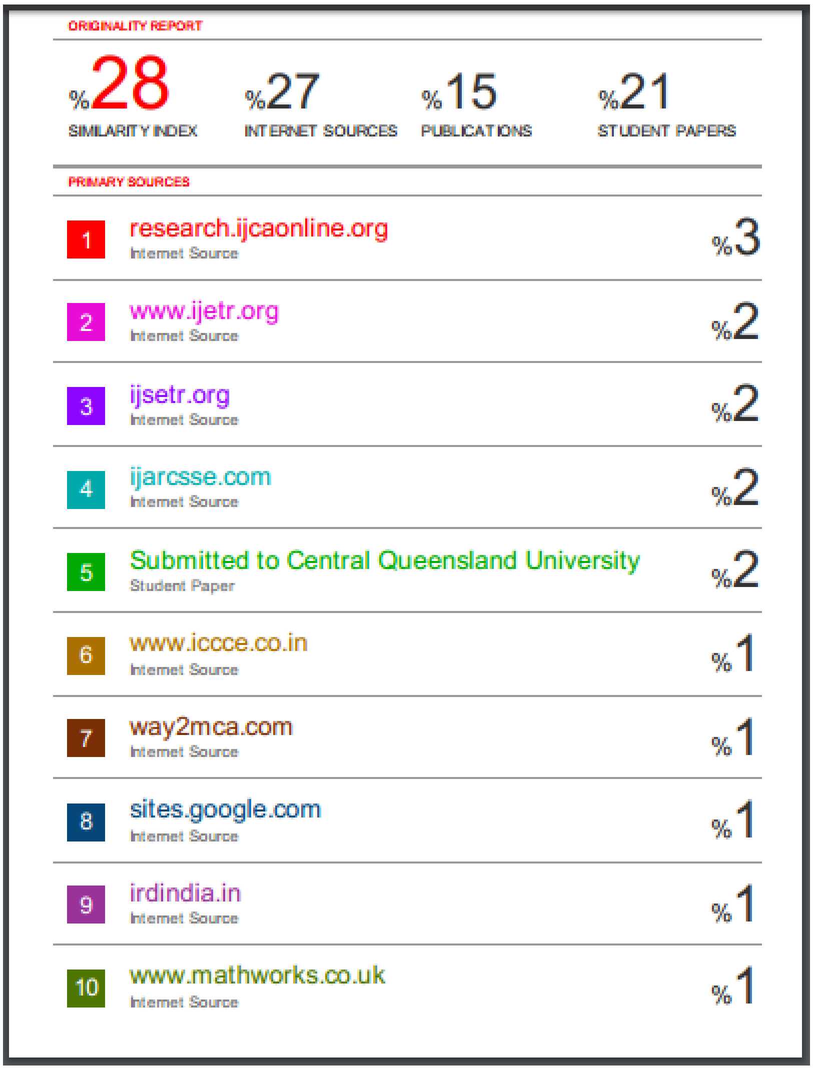
Certificate

This is to certify that Bro/Sis. Shraddha Bhanushali *has participated in Conference based on theme "Artificial Intelligence: Innovation towards future" in the category* Artificial Neural Network *in Prakalpa'17 held on 31st March, 2017*

Principal

Faculty Advisor

Plagiarism report for final BE project report.

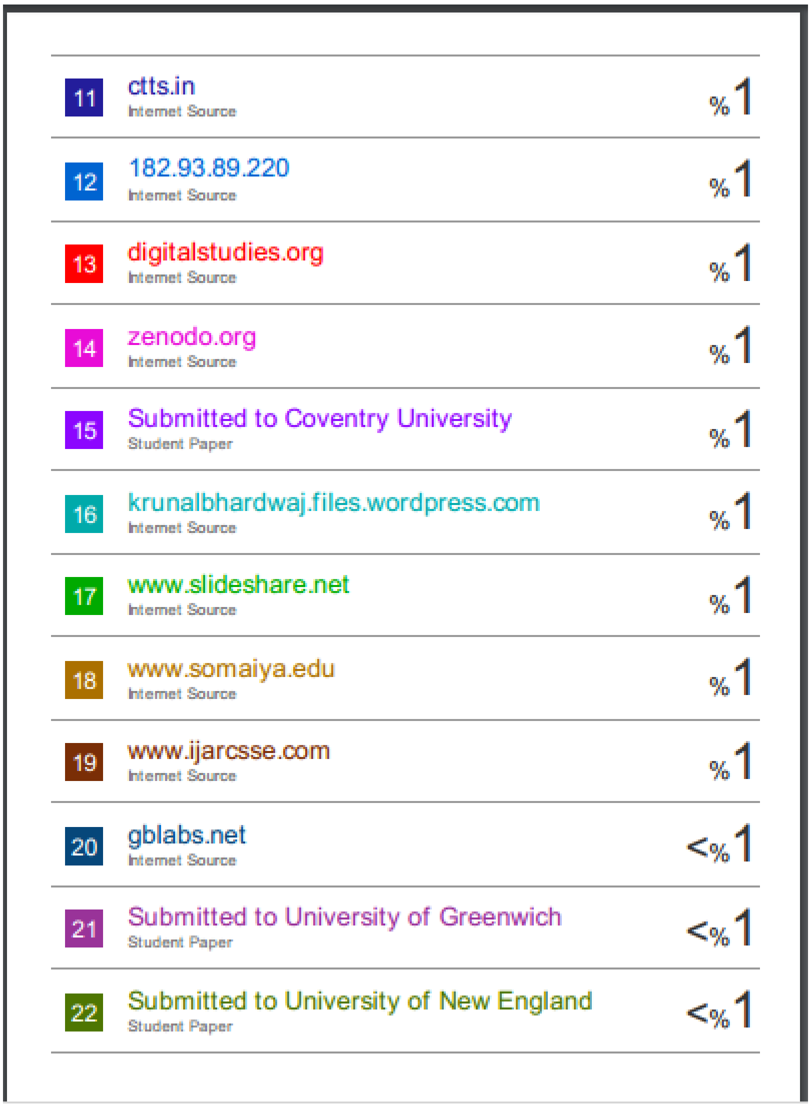


ORIGINALITY REPORT

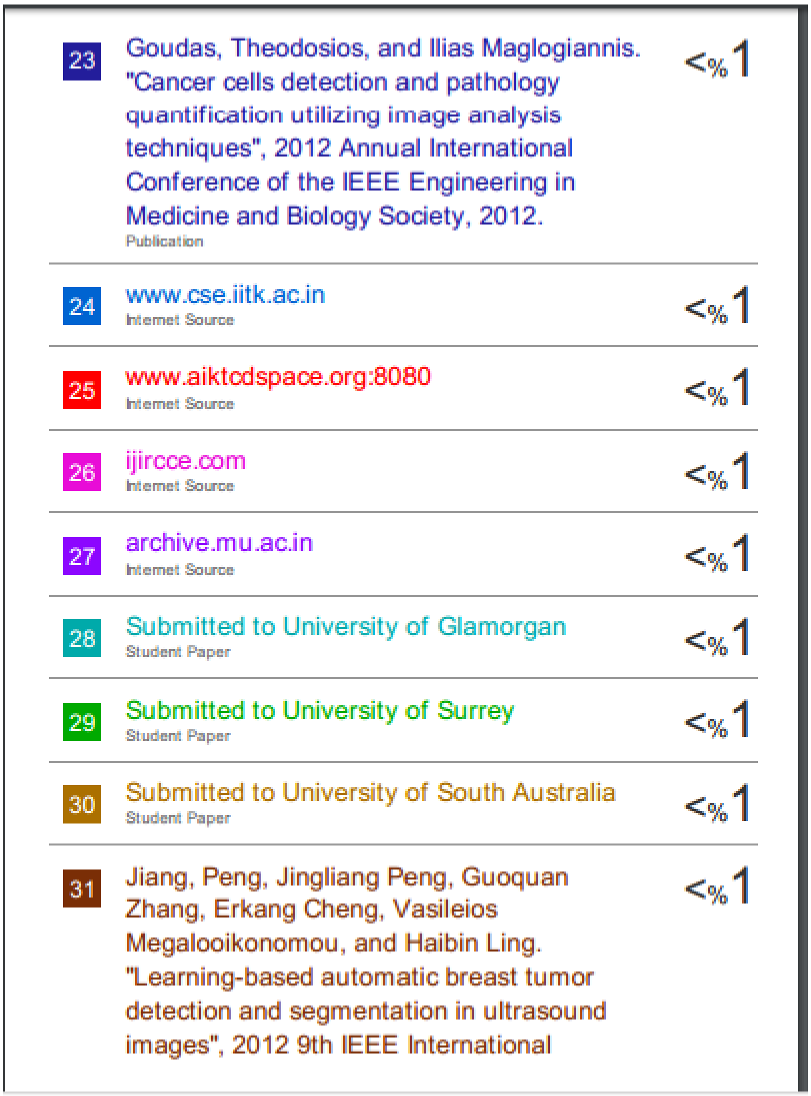
%28	%27	%15	%21
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

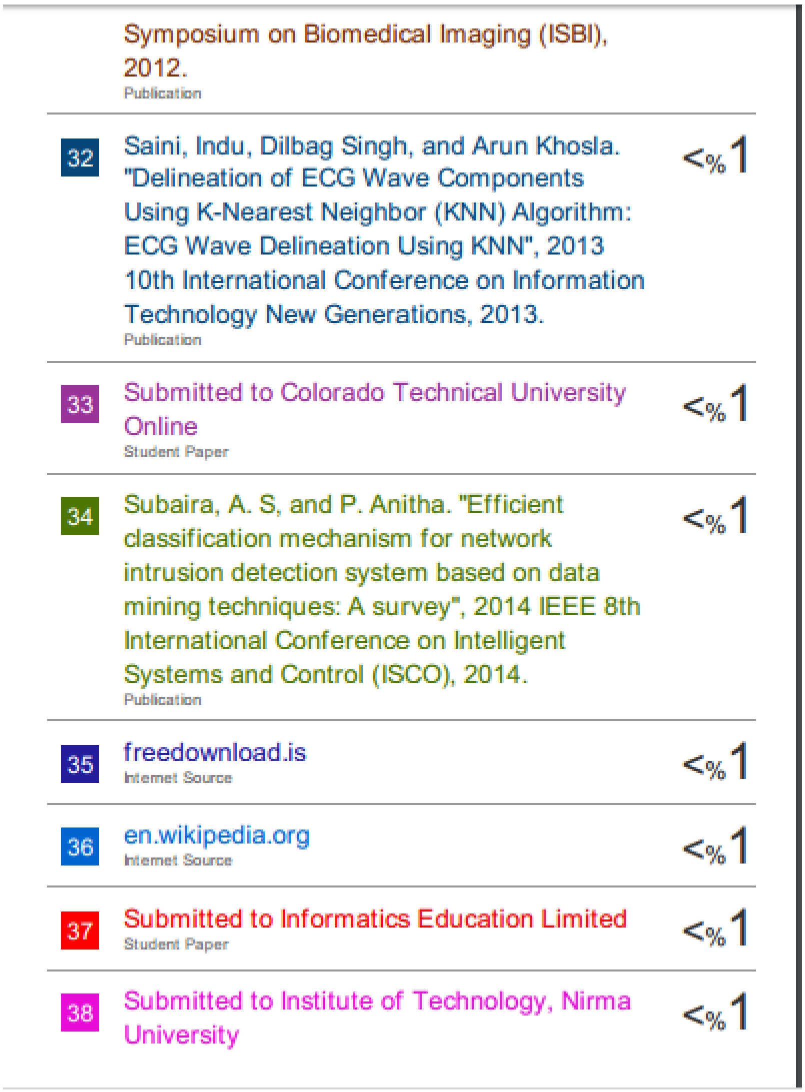
1	research.ijcaonline.org Internet Source	%3
2	www.ijetr.org Internet Source	%2
3	ijsetr.org Internet Source	%2
4	ijarcsse.com Internet Source	%2
5	Submitted to Central Queensland University Student Paper	%2
6	www.iccce.co.in Internet Source	%1
7	way2mca.com Internet Source	%1
8	sites.google.com Internet Source	%1
9	irdindia.in Internet Source	%1
10	www.mathworks.co.uk Internet Source	%1



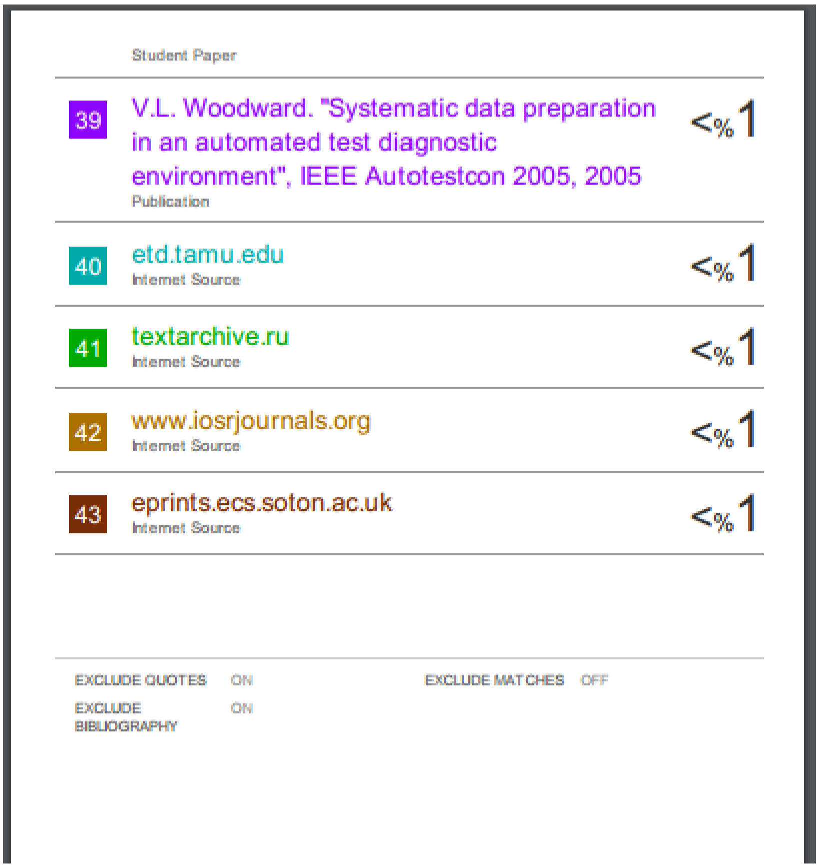
11	ctts.in Internet Source	%1
12	182.93.89.220 Internet Source	%1
13	digitalstudies.org Internet Source	%1
14	zenodo.org Internet Source	%1
15	Submitted to Coventry University Student Paper	%1
16	krunalbhardwaj.files.wordpress.com Internet Source	%1
17	www.slideshare.net Internet Source	%1
18	www.somaiya.edu Internet Source	%1
19	www.ijarcsse.com Internet Source	%1
20	gblabs.net Internet Source	<%1
21	Submitted to University of Greenwich Student Paper	<%1
22	Submitted to University of New England Student Paper	<%1



23	Goudas, Theodosios, and Ilias Maglogiannis. "Cancer cells detection and pathology quantification utilizing image analysis techniques", 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012. Publication	<%1
24	www.cse.iitk.ac.in Internet Source	<%1
25	www.aiktcdspace.org:8080 Internet Source	<%1
26	ijircce.com Internet Source	<%1
27	archive.mu.ac.in Internet Source	<%1
28	Submitted to University of Glamorgan Student Paper	<%1
29	Submitted to University of Surrey Student Paper	<%1
30	Submitted to University of South Australia Student Paper	<%1
31	Jiang, Peng, Jingliang Peng, Guoquan Zhang, Erkang Cheng, Vasileios Megalooikonomou, and Haibin Ling. "Learning-based automatic breast tumor detection and segmentation in ultrasound images", 2012 9th IEEE International	<%1



	Symposium on Biomedical Imaging (ISBI), 2012. Publication	
32	Saini, Indu, Dilbag Singh, and Arun Khosla. "Delineation of ECG Wave Components Using K-Nearest Neighbor (KNN) Algorithm: ECG Wave Delineation Using KNN", 2013 10th International Conference on Information Technology New Generations, 2013. Publication	<%1
33	Submitted to Colorado Technical University Online Student Paper	<%1
34	Subaira, A. S, and P. Anitha. "Efficient classification mechanism for network intrusion detection system based on data mining techniques: A survey", 2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO), 2014. Publication	<%1
35	freedownload.is Internet Source	<%1
36	en.wikipedia.org Internet Source	<%1
37	Submitted to Informatics Education Limited Student Paper	<%1
38	Submitted to Institute of Technology, Nirma University	<%1



Student Paper

39	V.L. Woodward. "Systematic data preparation in an automated test diagnostic environment", IEEE Autotestcon 2005, 2005 Publication	<%1
40	etd.tamu.edu Internet Source	<%1
41	textarchive.ru Internet Source	<%1
42	www.iosrjournals.org Internet Source	<%1
43	eprints.ecs.soton.ac.uk Internet Source	<%1

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		

Head,

Department of Computer Engineering