An Automatic Image Enhancement and Analysis Technique for Head and Neck Cancer Detection

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Pooja Gupta
(Roll No. 801632036)

Under the supervision of:
Dr. Avleen Kaur Malhi
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

June 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*An Automatic Image Enhancement and Analysis Technique for Head and Neck Cancer Detection.*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Avleen Kaur Malhi* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Pooja Gupta

Signature:

Pooja Gupta

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

9/6/17

Dr. Avleen Kaur Malhi

Assistant Professor

Computer Science and Engineering Department

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

(Dr. S S Bhatia)
Dean (Academic Affairs)
Thapar Institute of Engineering and Technology
Patiala

ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the guidance, assistance and help I have received from my supervisor, Dr. Avleen Kaur Malhi. I am thankful for her continual support, encouragement, and invaluable suggestions. She not only provided me help whenever needed, but also the resources required to complete this thesis report on time.

I am also thankful to Dr. Maninder Singh for his kind help and cooperation.

I would also like to thank all the staff members of Computer Science and Engineering Department for providing me all the facilities required for the completion of my thesis work.

I would like to say thanks for support of my classmates. I want to express my appreciation to every person who contributed with either inspirational or actual work to this thesis.

I am highly grateful to my parents and brother for the inspiration and ever encouraging moral support, which enabled me to pursue my studies.

Pooja Gupta

ABSTRACT

Every year, thousands of people are diagnosed with head and neck cancer. In hospitals, head and neck cancer detection is done by radiation therapy but it has some side effects and due to this therapy life quality of patient becomes less. It is detected manually by taking Computed Tomography images but it is very time consuming method. Therefore aim of this research is to detect cancer using some machine learning algorithms as this cancer rapidly increases nowadays. Head and neck cancer detection is performed by collecting 26019 CT scan images from Cancer Imaging Archive (TCIA). This research mainly focuses on classifier deep learning framework in h2o and decision tree followed by ensembling of both which gives better accuracy. Firstly, CT scan image of head and neck cancer is given as input to the system and processed through the image processing technique called weiner filter. Then the images are processed through the segmentation technique called fuzzy c-means algorithm. After that, feature extraction technique named Gray Level Co-Occurrence Matrix (GLCM) is used to extract the features. These features are given to classifier to train the model and finally it obtains the satisfactory results with 99.41% accuracy.

Index Terms- Head and neck cancer, Computed Tomography, Deep learning framework, Gray Level Co-Occurrence Matrix (GLCM).

TABLE OF CONTENTS

Certificate i

Acknowledgement ii

Abstract iii

Table of Contents iv

List of Figures vii

List of Tables ix

List of Abbrevations x

Chapter 1 Introduction **1**

1.1 Types of head and neck cancer 2

1.2 Imaging Techniques 3

1.2.1 Computed Tomography 4

1.3 Objective of Thesis 5

1.4 Importance 5

1.5 Structure of Thesis 5

Chapter 2 Literature Survey **7**

2.1 Head and neck cancer 7

2.2 Management of Head and Neck cancer 8

2.3 Review on Comprehensive Survey on Preprocessing Techniques 9

2.3.1 Image Denoising 9

2.3.2 Medical Images 9

2.3.3 Dicom Images 10

2.3.4 CT scan Images 10

2.4 Segmentation Approaches for Head and neck Cancer 10

2.5 Feature Extraction Techniques for Medical images 11

2.6 Machine Learning Approaches for diagnosis of Cancer 11

2.6.1 Classifiers 11

2.6.2 Ensemble 13

2.7 Comparitive Study 13

Chapter 3 Problem Statement **17**

3.1 Problem Definition 17

3.2 Gap Analysis 17

Chapter 4 Methodology **18**

4.1 Proposed Mechanism 19

4.2 Image Preprocessing 20

4.1.1 Techniques To Reduce Noise In Image 21

4.1.2 Comparison Parameters 23

4.2 Segmentation 24

4.2.1 Fuzzy c-means clustering 25

4.3 Feature Extraction 26

4.4 Classification 29

4.4.1 Deep learning in h2o 29

4.4.2 Decision Tree 31

4.4.3 Performance Metrics for Classification Problems 32

4.5 Ensemble of models 35

4.5.1 Types of ensembling 36

Chapter 5 Result Analysis 38

5.1 Deep learning in h2o 42

5.2 Decision Tree 44

5.3 Ensembled Results 46

5.4 Comparitive Analysis 46

Chapter 6 Conclusion and Future Scope 48

6.1 Conclusion 48

6.2 Future Work 49

Refernces 50

Publications 54

Plagiarism Report 55

LIST OF FIGURES

Figure 1.1 Different locations of head and neck cancer..... …….. 1

Figure 1.2 Division of head and neck cancer squamous cell carcinoma 2

Figure 1.3 Thyroid Cancer 3

Figure 1.4 Salivary Gland 3

Figure 1.5 Head and neck cancer sample 4

Figure 2.1 Cancers in India for both sexes 7

Figure 2.2 Comparison between filters 10

Figure 4.1 WorkFlow 18

Figure 4.2 Proposed Mechanism 20

Figure 4.3 Filtering Techniques 21

Figure 4.4 Clustering Techniques 24

Figure 4.5 Flowchart of Fuzzy c means Algorithm 25

Figure 4.6 Construction of Gray Level Co-occurrence Matrix 26

Figure 4.7 Different Orientation of pixels 27

Figure 4.8 Deep learning neural network 30

Figure 4.9 Multilayer Feed Forward Neural Network 31

Figure 4.10 Basic Terminologies of Decision Tree 32

Figure 4.11 Accuracy Representation 34

Figure 4.12 Precision Representation 34

Figure 4.13 Recall Representation 35

Figure 4.14 Specificity Representation 35

Figure 4.15 Ensembling of Models 36

Figure 5.1 Image 1 – Results of median and weiner filter o/p 38

Figure 5.2 Image 1 – Results of Average and Guassian filter o/p 39

Figure 5.3 Graphical Representation of filtering techniques 40

Figure 5.4 Original, Preprocessed and Segmented Images 40

Figure 5.5 Histogram of Class 42

Figure 5.6 ROC curve 43

Figure 5.7 Decision Tree 44

Figure 5.8 Graphical Representation of different Classifier 47

LIST OF TABLES

Table 2.1 Relative proportion of head and neck cancer patients 9

Table 2.2 Literature Review in Table Format 13

Table 4.1 Confusion Matrix 33

Table 4.2 Averaging 36

Table 4.3 Majority vote 36

Table 4.5 Weighted Average 37

Table 5.1 Comparison Parameters for image 1 39

Table 5.2 First order statistical feature of few images 40

Table 5.3 Second order statistical feature of few images 40

Table 5.4 Confusion Matrix of Deep learning in h2o for Train Data 42

Table 5.5 Confusion Matrix of Deep learning in h2o for Test Data 43

Table 5.6 Model Performance Parameters of Deep learning in h2o for Train Data 43

Table 5.7 Model Performance Parameters of Deep learning in h2o for Test Data 43

Table 5.8 Confusion Matrix of Decision Tree for Train Data 45

Table 5.9 Confusion Matrix of Decision Tree for Test Data 45

Table 5.10 Model Performance Parameters of Decision Tree for Train Data 45

Table 5.11 Model Performance Parameters of Decision Tree for Test Data 45

Table 5.12 Ensembled Result 48

Table 5.13 Comparison of different Classifiers based on accuracy 48

LIST OF ABBREVATIONS

1. TCIA: The Cancer Imaging Archive
2. HPV: Human papillomavirus
3. HNSCC: Head and neck squamous cell carcinoma
4. CT: Computed Tomography
5. PET: Positron Emission Tomography
6. MRI: Magnetic Resonance Imaging

CHAPTER 1

INTRODUCTION

Head and neck cancer is a group of cancers of mouth, sinuses, nose or throat and it accounts for about 3% of all cancers in United States. In 2015, 59340 people suffered from head and neck cancer and from that about 12290 deaths occurred. Therefore head and neck cancer should be taken seriously as this cancer has major consequences. The most common treatment for head and neck cancer are radiation therapy and surgery. Head and cancer can occur in any organ in the region of head and neck cancer. There are possibly 30 different organs where head and neck cancer can develop. Figure 1.1 shows the different locations where head and neck cancer can occur.

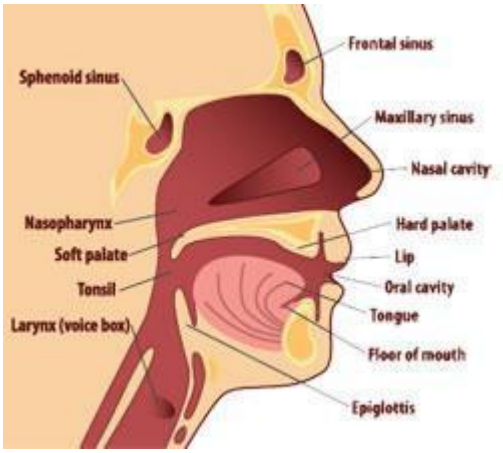


Figure 1.1: Different locations where head and neck cancer can occur in human body [1]

There are various risk factors for head and neck cancer from which alcohol consumption and use of tobacco mainly increases chances of head and neck cancer. Other risk factors are:

1. Smoking and Drinking
2. Infection from wood or nickel
3. Due to lack of nutrition
4. Weak immune system

5. Infection from viruses likes Human papillomavirus (HPV) - It infects the oropharynx that includes tonsils and back of throat and this is infected by sexual transmission.

1.1 Types of Head and Neck Cancer

Head and neck cancer is basically of two types:

1. **Head and neck squamous cell carcinoma (HNSCC)** – The division of head and neck cancer squamous cell carcinoma (HNSCC) parts is shown in figure 2. It includes:

Oral Cavity – Cancer that occurs in the mouth is called oral cavity

Throat (pharynx) – This cancer affects the following parts of pharynx

- Nasopharynx – Near nasal cavity, if cancer is in upper layer of throat.
- Oropharynx – Near mouth, if cancer is in middle layer of throat.
- Hypopharynx – Near voice box, if cancer is in lower layer of throat

Voice box (larynx) – The voice box, additionally called the larynx, is a short path that associates the lower layer of the throat (hypopharynx) with the windpipe (trachea).

Nasal cavity and paranasal sinuses - The nasal cavity is the huge, empty space inside the nose.

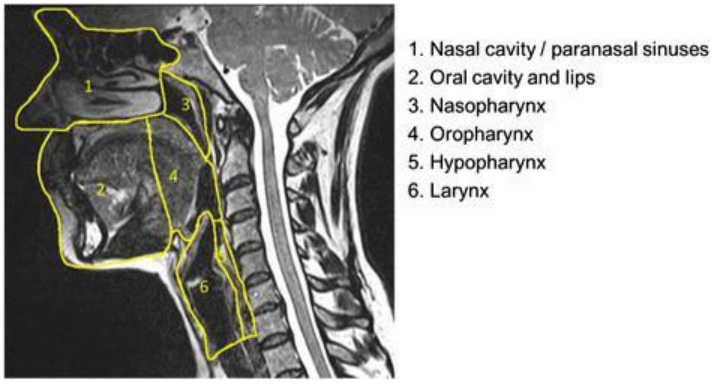


Figure 1.2: Division of head and neck cancer squamous cell carcinoma [2].

2. **Salivary Gland and Thyroid Cancer**

Thyroid Cancer - Thyroid cancer is mainly occurred in women than men and present in front of neck and thyroid cancer has mainly two types. Papillary and

Follicular cancers both occur in follicular cells and growing slowly but Papillary spreads to the lymph nodes. Figure 1.3 shows thyroid cancer.

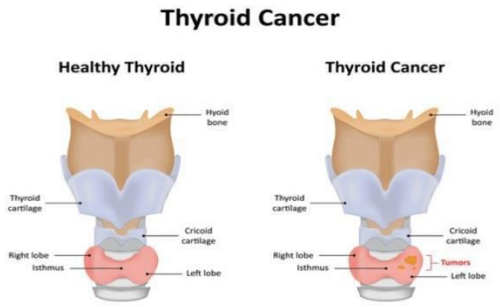


Figure 1.3: Thyroid Cancer [3].

Salivary Gland – By this, mouth keeps moist and helps to slide down food into stomach. There are three types of salivary glands. Figure 1.4 shows types of salivary gland.

- Parotid gland – These glands are largest and present at both sides in front of ear.
- Sublingual gland – These glands are present underneath the tongue.
- Submandibular gland - These glands are present underneath the jawbone.

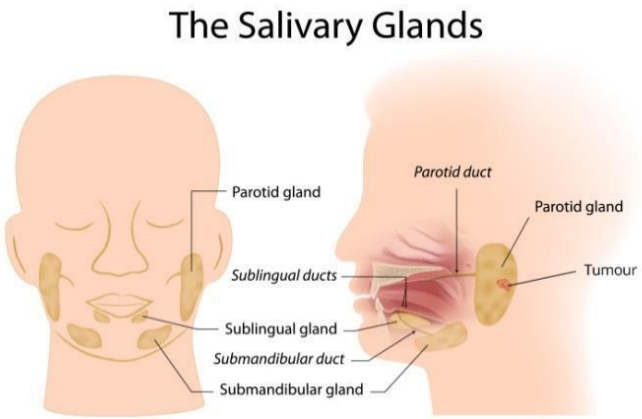


Figure 1.4: Salivary Glands [3]

1.2 Imaging Techniques

Various types of imaging techniques are used to detect head and neck cancer like Computed Tomography (CT), Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) [4]. In this work, CT scan images have been used.

1.2.1 Computed Tomography – CT scan images are used for diagnosis of cancer in patient. By using CT scan we obtain the information about internal structure of body. To obtain better results on the image, contrast medium is given to the patient in the form of liquid or through injection so that image has more clarity. CT scan images are easily evaluated than MRI images. Figure 1.5 shows CT scan image sample of head and neck cancer.

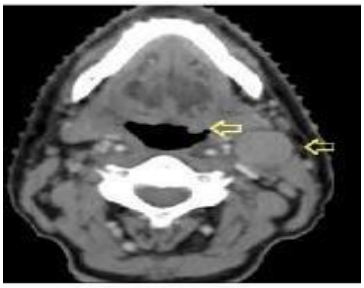


Figure 1.5: Head and neck cancer sample [5]

Currently, physicians have been manually detecting the tumor location from CT scan images which is quite time consuming and frustrating. This method is inefficient and also leads to inaccuracy. Therefore there is requirement for automating the process so that the tumour location can be automatically detected from images with machine learning algorithms without the actual presence of any physician. Depending on this, there is a need to develop a framework which can do classification based segmentation based on head and neck cancer images.

The objective of the thesis is to automatically detect the locations of the head and neck cancer from the collected CT scan images. The Image data set has been collected from The Cancer Imaging Archive (TCIA) where about 26019 images have been collected. Firstly, the images are preprocessed and then the image segmentation is done to improve the image visually which will be easy to analyze. After the segmented images are obtained, the feature extraction is done to extract the required features which can be used in image classification and tumour location detection. Finally deep learning with H2O and decision tree is used to model training and testing and ensembling is done to achieve better accuracy and it is also compared with other classifiers for better accuracy.

1.3 Objectives

The main objective of thesis can be summarized as:

1. To study, explore and analyse already existing methods to detect head and neck cancer and overcome the limitations of that existing methods with new approaches.
2. To propose image analysis framework that can be easily used for head and neck cancer prediction of CT scan images and gives accurate results.
3. To detect head and neck cancer by using feature extraction and then classifies them by doing ensemble of classifier models
4. To test and validate the proposed technique.

1.4 Importance

Achieving higher accuracy in prediction of head and neck cancer is very crucial task. Head and neck cancer has various different organs and therefore it is important to consider all the organs for treatment. It is also important to identify cancer as soon as possible so that patient can start their treatment early. Detecting cancer is one of the major tasks. Therefore if that technique used for cancer prediction that gives high accuracy then it will be beneficial for all cancer patients and also helpful for doctors. It will also give high sensitivity and high specificity by which it will be easier to predict cancer. There are very few techniques of machine learning used in this area to overcome limitation of manually detection of cancer so by comparing all techniques will help in estimating the best technique among them for cancer prediction.

1.5 Structure of Thesis

The thesis is grouped into 6 chapters including literature review, problem statement, methodology, experimental results, summary, conclusion and future scope followed by references.

Chapter 1 provides the subject area in which thesis work has been done and also provides the objective of the thesis.

Chapter 2 provides the literature survey of all the different methodologies to detect head and neck cancer that have been used so far.

Chapter 3 discusses about the problem statement, gap analysis

Chapter 4 provides the methodology that means how the problem has been solved which includes preprocessing, segmentation, feature extraction and ensembling. Chapter 5 includes the results analysis and discussions.

Chapter 6 presents summary, conclusion and also future scope which discusses what further can be done in this research area.

CHAPTER 2

LITERATURE SURVEY

Keeping in mind about head and neck cancer it is important to predict as advancement in machine learning allows us to develop image recognition tools which help in this process by taking less time and gives more accurate results than manual treatment.

2.1 Head and Neck Cancer

In 2003, head and neck cancer was found in 37000 men and women [6]. Head and neck cancer becomes major health problem in India as 57.5% of people in India suffer from this problem which affects their life style [7]. Head and neck cancer can emerge in any part of head and neck. This cancer starts from primary location and spread in various organs and this process is called metastasis. Primary location is the location from where the tumor starts growing. This cancer can be spread through lymph nodes or through blood vessels but the common way is through lymph nodes. Head and neck cancer [8] is a group of cancers of the mouth, sinuses, nose or throat and it accounts for about 3% of all cancers in the United States. The most common treatments for head and neck cancer are radiation therapy and surgery.

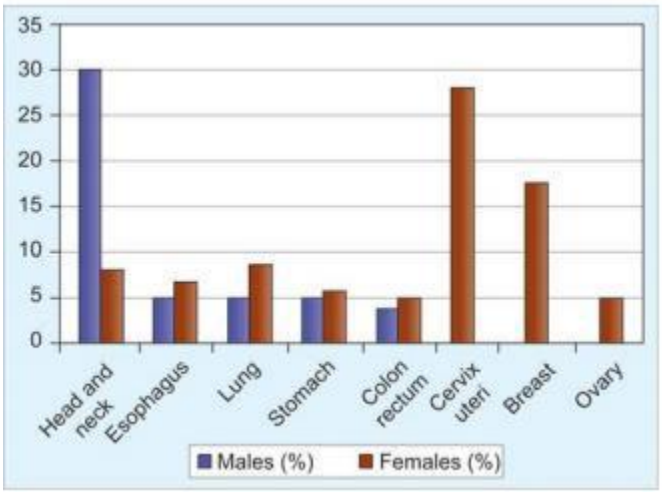


Figure 2.1: Cancers in India for both sexes [7]

In 2015, 37900 deaths [9] were occurred due to head and neck cancer. Head and neck has various types of cancer. Therefore it is necessary to determine which type of cancer a patient have so that a patient gets right treatment. Sometimes by taking improper treatment cancer comes back. This is due to some cancer cells remain in that part where cancer occurs which cannot be even recognized by scans. Therefore there is no guarantee for any treatment. If cancer can again occur in same location then it is called local recurrence. If it occurs in different location then it is called metastasis [10]. Head and neck cancer can be diagnosed by certain tests. For this doctor first examine the patient and then accordingly certain tests are performed. Test can be ultrasound scan or nasendoscopy or examination under anesthetic (EUA) or a trans-nasal flexible laryngo-oesophagoscopy (TNFLO) or a biopsy. Head and neck cancer can occur due to various reasons. Generally cancer will develop when person is over 65 and older. Other reasons for cancer are changing life style, age. But sometimes these reasons are not responsible for developing cancer this happens for no reason.

2.2 Management of Head and Neck cancer

Danish head and neck cancer (DAHANCA) group arranged the head and neck cancer management in Denmark. This group guarantees about the investigating and treatment of cancer by collecting clinical data. This group is also associated with other organization and together they ensures about the quality of treatment. Firstly, always go to specialist for consultation and then they suggest what type of treatment a person need. As soon as possible start the treatment otherwise it will spread to other parts of body. There are no rules for head and neck cancer management. Table 2.1 shows surgery with radiotherapy given in Mumbai is more than other cities [11].

Table 2.1: Relative proportion of head and neck cancer patients

Type of Treatment	Mumbai	Bengaluru	Chennai	Trivandrm	Dibrugarh
Surgery	19.7%	11%	4.2%	7.2%	3.3%
Radiotherapy	18.8	54.8	44.7	44.6	85.9
Chemotherapy	6.1	5.2	1.2	5.4	2.3
Surgery + Radiotherapy	36.9	16	17.9	9.9	4.4
Surgery + Chemotherapy	0.8	0.8	0	0.8	0.2
Radiotherapy + Chemotherapy	15.5	10.6	28.2	26.2	3.4

2.3 Review on Preprocessing Techniques

Preprocessing is necessary for better quality of image and we can use various preprocessing techniques according to the type of image. This section is further divided into various sections to represent which technique was used by authors for particular type of image.

2.3.1 Image Denoising

Nagu *et al*. [12] proposed the technique where they use median filter and weiner filter for removal of noise from image. This paper concluded that after filtration new image is almost same as original image. Therefore it is better to use adaptive median filter rather than median filter.

2.3.2 Medical Images

Perumal *et al*. [13] proposed the technique where median, weiner and Guassian filter is used for removal of noise from medical images. Here difference is found out by quality

and pattern of noise. This paper concluded that median filter is better than weiner and Guassian filter.

2.3.3 Dicom Images

Labeeb *et al.* [14] proposed the technique where they compared three filters median, weiner 5*5 and weiner 3*3 for dicom images and concluded that weiner 3*3 performed better than other two. The figure 2.2 shows the comparison of filters.

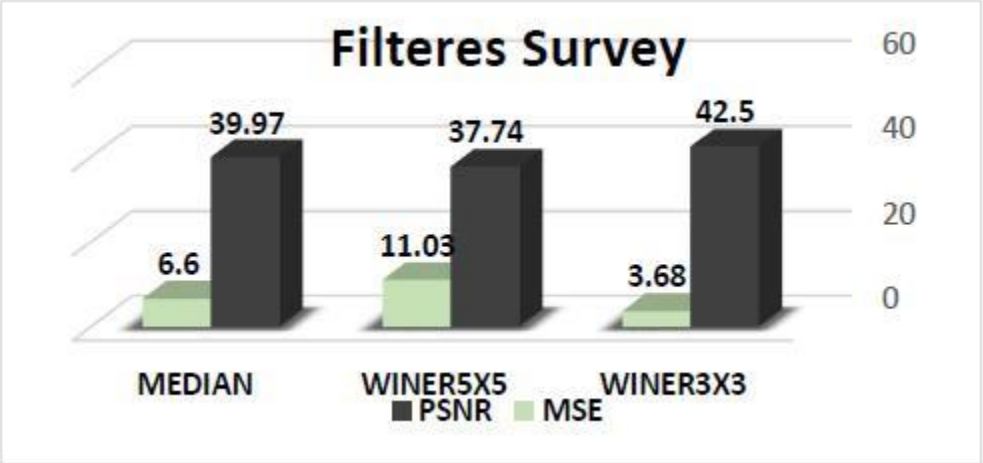


Figure 2.2: Comparison between filters [14]

2.3.4 CT scan images

Chhabra *et al*. [15] proposed a technique for CT scan images where they compare wavelet decomposition, weiner filter, median filter, NLM filter, Anisotropic Diffusion, Wave atom Transform, Anisotropic Diffusion in Wavelet Domain using evaluation paremeters and concluded that Anisotropic Diffusion in Wavelet Domain is better for CT scan images.

2.4 Segmentation Approaches for Head and neck Cancer

Kaur *et al.* [16] compared various segmentation techniques that are used for object detection and for medical images but it difficult to recognize the whole image therefore image is partitioned into segments and concluded that all segmentation techniques are useful for medical images but also concluded that not every image is suitable for particular segmentation method.

Norouzi *et al.* [17] gives the details about various segmentation methods and describes the latest method with their pros and cons. Here every segmentation method is described with features along with their advantages in Computed Tomography images and concluded that fuzzy c means is better than k means in clustering technique for medical images.

2.5 Feature Extraction Techniques for Medical images

Ahmad *et al.* [18] uses six feature extraction techniques and compares the feature extraction techniques of CT scan images on content based image retrieval. This method also helps in extracting the disease that is of similar kind and concluded that Gray Level histogram is best for intensity, Discrete Wavelet Frame is best for texture and Fourier Descriptor is best for shape.

Singh *et al.* [19] covers the importance of feature extraction from segmented image. Here in this paper histogram equalization followed by thresholding and then some random features are extracted that gives important information and this is essential for achieving high accuracy and also reduces the complexity.

2.6 Machine Learning Approaches for diagnosis of Cancer

The various machine learning approaches have been used so far to detect and classify various types of cancers. The various methods have been studied in this section.

2.6.1 Classifiers

Since decades, different flavours of CNN have been used for classification and detection of cancer and other objects. It was Introduced in 1980 [20] and gradually improved over the next two decades [21][22][23]. This section discusses the literature related to the cancer detection deep learning approaches which have already been proposed in literature. Fakoor *et al.* [24] proposed a deep learning approach for cancer detection where the authors used unsupervised feature learning. It helps to apply the data for detection of multiple types of cancers for automatic feature formation for analysis and detection of specific type of cancer outperforming the other techniques. Another method was proposed by Albarqouni *et al*. [25] for detection of mitosis in breast cancer from

histology images. The authors used the convolutional neural network for handling the data aggregations of cancer images by using a crowdsourcing layer. The experimental analysis was performed on learning from crowds for training CNNs for data collected from crowds. Thus, the paper provides the CNN functionality for necessity of data aggregation in breast cancer histology images. A new deep learning based framework for cancer detection was proposed by Cruz-Roa *et al*. [26] for automated basal cell carcinoma cancer detection. It highlighted the visual patterns for discrimination between normal and cancerous tissues and spotlighted the image regions for diagnosis purposes. The experiments were performed on 1417 images with obtained F-measure as 89.4% and accuracy of 91.4%.

Sirinukunwattana *et al.* [27] proposed a deep learning based model for classification and detection of colon cancer from histology images. The nucleus detection was performed with the help of SC-CNN and Neighbouring Ensemble Predictor was used for the prediction of detected cell nuclei's class label which helped in producing the highest F1 scores compared to other methods. Mitosis detection in breast histology images is done with the help of max pool convolutional networks as proposed by Ciresan *et al.* [28]. The training of the network was done in such a manner that each pixel was classified for patch centered on pixel. Deep learning approach was proposed by Cruz *et al*. [29] for breast cancer detection to automate the process of identification of invasive tumour from whole slide images. The approach was evaluated on more than 200 cases from Cancer genome yielding a better accuracy of detection of cancer. An approach for prostate cancer detection was proposed by Tsehay *et al*. [30] where he used convolutional neural networks to automate the process of detecting lesions on mpMRI. It yielded a detection rate of 80% with false positive rate of 20% which was also translated into 94% detection rate with per patient false positives of 10. Also different classifiers of decision tree have been used in classification and detection of cancer and other objects. This section discusses the literature related to the cancer detection decision tree classifiers. Sujatha *et al.* [31] proposed the method where performance evaluation of decision tree classifiers ID3, C4.5 and CART are analysed on tumour datasets. It helps to achieve better accuracy and also concluded that C4.5 is better than others. An approach for breast cancer detection was proposed by A.Elsayad [32] where he used four different decision tree

algorithms C&R, CHAID, QUEST, and C5.0 to detect cancer where dataset is partitioned into 70% and 30% data. Hamsagayathri *et al.* [33] proposed a method where he used priority based decision classifier for detection of breast cancer where they achieved accuracy of 98.51%, sensitivity 91.15% and specificity 99.86%.

2.6.2 Ensemble

Hijazi *et al.* [34] compares the accuracy of single classifiers with the ensemble of classifiers He gave the conclusion that ensemble classifiers are better than single classifiers and ensembling model is better for detecting cancer. T. G. Dietterich *et al.* [35] proposed the various ensembling methods like error-correcting output coding, Bagging, and boosting and then compared these methods and concluded that ensembling gives better results than individual models. A. Onan [36] proposed the performance of ensembling for breast cancer and also concluded that Random Subspace, Dagging and Multi Boosting are ensembling methods for medical data.

2.7 Comparative Study

The comparative study of the proposed techniques has been done and in shown in table 2.2.

Table 2.2: Comparative study of related literature

Authors	**Description**	**Purpose**	**Results**
Nagu *et al.* [12]	Used median filter and weiner filter for removal of noise from image	Image De-noising	Adaptive Median Filter is better
Perumal1 *et al.* [13]	Used median, weiner and Guassian filter	Image De-noising from medical images	Median Filter is better
Labeeb *et al.* [14]	Used median, weiner 5*5 and	Image De-noising	Weiner 3*3 filter is

	weiner 3*3 filter	from dicom images	better
Chhabra *et al.* [15]	Wavelet decomposition, weiner filter, median filter, NLM filter, Anisotropic Diffusion, Wave atom Transform, Anisotropic Diffusion in Wavelet Domain	Image De-noising from CT scan images	Anisotropic Diffusion in Wavelet Domain is better
Norouzi *et al.* [17]	Used various segmentation techniques.	Segmentation	Fuzzy c means is better than k means in clustering technique for medical images.
Ahmad *et al*.[18]	Compares the feature extraction techniques of CT scan images on content based.	Feature Extraction from medical images	Gray Level histogram is best for intensity, Discrete Wavelet Frame is best for texture and Fourier Descriptor is best for shape
Singh *et al*. [19]	Describes the importance of feature extraction from segmented	For achieving high accuracy	NA

	image		
Cruz-Roa *et al.* [26]	It highlighted the visual patterns for discrimination between normal and cancerous tissues and spotlighted the image regions for diagnosis purposes.	Deep learning approach for automated basal cell carcinoma cancer detection	F-measure as 89.4% and accuracy of 91.4%.
Tsehay *et al.* [30]	He used convolutional neural networks to automate the process of detecting lesions on mpMRI	Prostate cancer detection	Detection rate of 80% with FPR of 20% which was also translated into 94% detection rate with per patient false positives of 10.
Sujatha *et al.* [31]	Performance evaluation of decision tree classifiers ID3, C4.5 andCART are analysed on tumour datasets	To achieve better accuracy	C4.5 is better
A.Elsayad [32]	Used C&R, CHAID, QUEST, C5.0 to detect	To detect breast cancer	C&R – 96.334% accuracy CHAID – 96.130%

	cancer		accuracy QUEST – 95.927% accuracy C5.0 - 97.963% accuracy
Hamsagayathri*et al.* [33]	Used priority based decision classifier	For detection of breast cancer	Accuracy of 98.51%, sensitivity 91.15% and specificity 99.86%.
Hijazi *et al.*[34]	Compared single classifiers with ensembled classifiers	Ensembling	Ensembled Classifier is better
T.G. Dietterich [35].	Used Bagging, Boosting and Coding	Ensemble the models	Ensembled model gives better result than individual results.
A. Onan [36]	Different ensembling algorithms are used	For detection of breast cancer	Random Subspace, Dagging and Multi Boosting are suitable for medical data

In existing approaches, no preprocessing has been done on images, as it is required for better image quality. Therefore, in proposed technique preprocessing has been done by comparing various preprocessing techniques and selected the one which gave better results. The proposed approach uses ensembling of deep learning in h2o and decision tree to achieve better accuracy.

CHAPTER 3

PROBLEM STATEMENT

3.1 Problem Definition

Head and neck cancer is a group of cancers of the mouth, sinuses, nose or throat and it accounts for about 3% of all cancers in the United States. The most common treatments for head and neck cancer are radiation therapy and surgery. Currently, physicians have been manually detecting the tumour location from CT scan images which is quite time consuming and frustrating. This method is inefficient and also leads to lesser accuracy. Therefore, there is requirement for automating the process so that the tumour location can automatically be detected from images with machine learning algorithms without the actual presence of any physician. Depending on this, there is a need to develop a framework which can do classification based on head and neck cancer images to get better accuracy and to achieve this accuracy, ensembling is done in the proposed work as ensembling calculates the overall prediction of the models and it is stronger than individual models.

3.2 Gap Analysis

The presently head and neck cancer detection techniques used less samples.

Accuracy achieved by these methods is less.

Pre-processing was not done in state-of-the-art techniques which are required for better image quality.

Achieving high sensitivity and specificity is required.

CHAPTER 4

METHODOLOGY

An automated method is proposed to detect head and neck cancer that only uses images of CT scan. CT scan images are taken from Cancer Imaging Archive (TCIA) which provides medical images. Total 26019 images are collected. Then, various image processing techniques are applied to detect cancer. Preprocessing is done to improve the quality of image. After doing preprocessing, segmentation is done for analyzing the image clearly. Then, feature extraction is done to extract the necessary parameters. Deep learning framework in h2o and decision tree is applied to these extracted features with 80% of the data from the dataset is used for train the system and 20% of the data from the dataset is used to test the system and then ensembling is done to get better results.

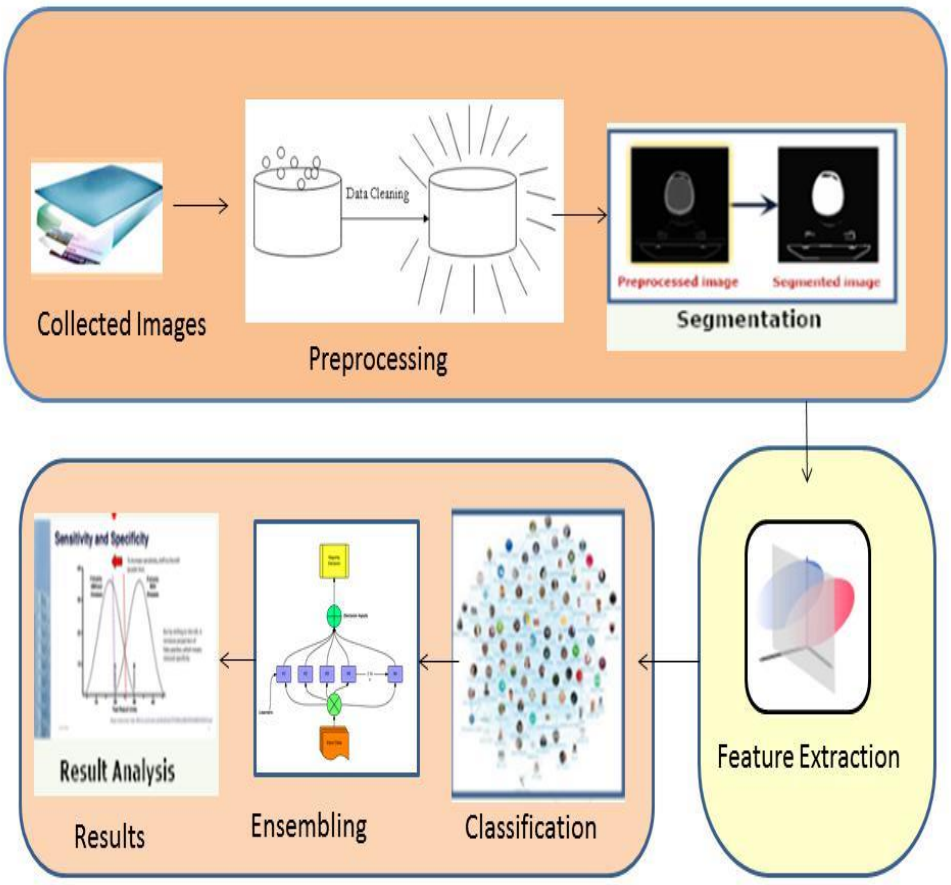


Figure 4.1: Workflow of the proposed work

This chapter covers the methodology followed to identify the cancer and classify the images of cancer. The subsequent sections explain the following steps: Section 4.1 covers the different pre-processing techniques to reduce noise and their comparison parameters in brief. Section 4.2 explains the segmentation method used for the pre-processed image. Section 4.3 explains the feature extraction technique used in research work. Last section, Section 4.4 focuses on the different classifiers and ensembling on those classifiers. Diagrammatic representation of methodology is shown in figure 4.1.

4.1 Proposed Approach

Here Machine Learning approach is used for detection of head and neck cancer. Proposed approach is shown in figure 4.2. Firstly images of head and neck cancer collection has been collected from Cancer Imaging Archive(TCIA) and then preprocessing section comes where filtering and cluster based segmentation has been done in MATLAB. To classify cancer there is a need to extract features from images and for that GLCM and histogram based features has been extracted. This has been also done in MATLAB. Now after feature extraction, to train the system deep learning in h2o and decision tree model has been applied and get the parameters named Confusion Matrix, Accuracy, Sensitivity, Specificity, Precision and ROC curve. Then ensembling has been done to achieve the best results and compare these results with the already existing methods.

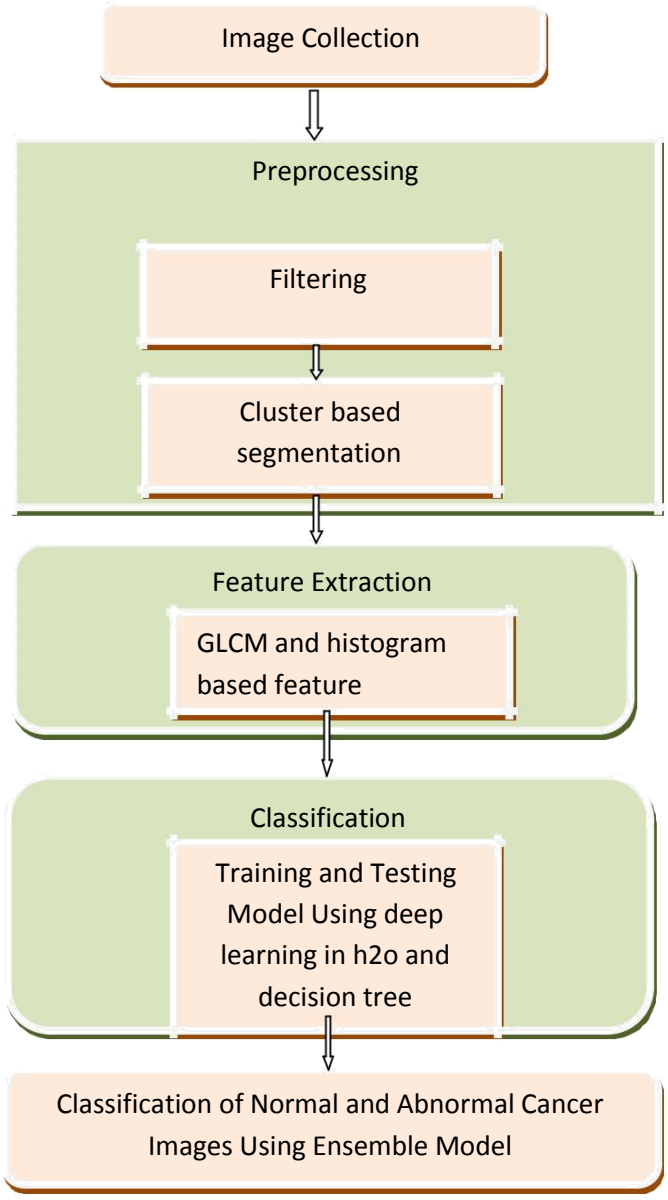


Figure 4.2: Proposed Approach for head and neck cancer detection

4.2 Image Preprocessing

Preprocessing means enhancement in image data that is used for further processing. As most of the images have noise that leads to poor quality of image hence unable to extract important information from that image. Therefore preprocessing is done to improve the quality of image. For preprocessing, various filtering techniques are used and got the best one by comparing them. Preprocessing is needed as CT (computed tomography) scan images are used and CT scan images generally impacted by technical parameters. One of

the parameter is radiation dose. Radiation dose increases the image quality but these radiations are harmful for humans and also lead to cancer. To avoid these radiations it is necessary to use another method that reduces noise in images and therefore filtering techniques comes into picture. Removal of noise plays a very important role in medical images for proper diagnosis of cancer. Various denoising algorithms are applied on images to get best result. The algorithms that are applied to images are Weiner Filter, Averaging Filter, Median Filter and Guassian Filter as shown in figure 4.3.

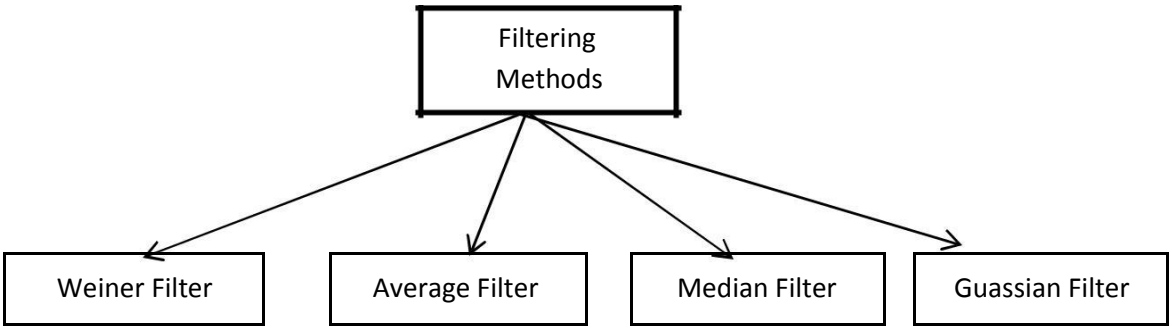


Figure 4.3: Filtering Techniques

4.2.1 Techniques to Reduce Noise in Images

Weiner Filter

Weiner filter is used to filter out the white noise and removes blurredness from the image. Throughout the image, it works as a fixed filter. It also takes less computation time to provide the solution. The output is not same as original image, but almost same. The performance function of weiner filter is as follows:

$$= [|( )|]^{2} \quad (1)$$

Where x is called the mean square error criterion as weiner filter optimized the overall mean square error.

Averaging Filter

Replacing each pixel value in an image by the average value of its neighbour is called averaging filter. It behaves as a convolution filter and based on a kernel. It removes the unwanted pixel values from their surroundings. Generally small

kernel like 3*3 is used but for more smoothing 5*5 kernel is used. An example of 5*5 single window of average filtering is given below:

Unfiltered Table

3	5	7	9	2
2	1	3	2	5
8	10	6	7	9
5	31	5	14	11
2	5	16	25	7

Filtered Table

3	5	7	9	2
2	1	3	2	5
8	10	6	7	9
5	31	5	14	11
2	5	16	25	7

Mean of unfiltered values is calculated as:

1. Firstly by taking sum of all unfiltered values
   =3+5+7+9+2+2+1+3+2+5+8+10+6+7+9+2+31+5+14+11+2+5+16+25+7=200
2. Mean= Sum of all unfiltered values / Total no. of unfiltered values = 200/25=8

Here center value of unfiltered table which is 6 is replaced by 8 in filtered table where 8 is the mean of all the twenty five values from unfiltered table.

Median Filter

There are two types of filter used for removing noise. Linear filter and Non-Linear Filter. By using linear filter, image gets blurred so to overcome this problem non-linear filter is used and Median Filter comes under the non-linear filter. In median filter centre values is replaced by median of all the neighbouring values and centre. Therefore they are good for medical images. An example of 3*3 window of median filtering is as follows:

Unfiltered table

3	5	7
2	1	3
8	10	6

Filtered Table

*	*	*
*	5	*
*	*	*

Sort the values of unfiltered table in increasing order as 1,2,3,3,5,6,7,8,10 Median value = 5. Here centre value of unfiltered table which is 1 is replaced by 5 in the filtered table which is the median of all the values from unfiltered table.

Guassian Filter

It is a type of linear filtering and faster than median filtering because in median filter first we have to sort the pixel values and it is time consuming. Guassian filter works using odd size of mask and mask value of point has been evaluated using Guassian function where Guassian function is as follows:

$$G(x) = \frac{1}{\sqrt{2\ ^2}} e^{\frac{x^2}{2\ ^2}} \quad (2)$$

Where is called standard deviation.

4.2.2 Comparison Parameters

Following are the comparison parameters to compare the results of the various filtering techniques.

Mean Square Error (MSE)

MSE is defined by calculating cumulative mean square error that occurred between compressed and original form of image. MSE measures the quality and fidelity of image. MSE is defined as follows:

$$MSE = \frac{1}{A * B} \sum_{m} \sum_{n} [M(m, n) - W(m, n)]^2 \quad (3)$$

where (M (m, n)) is the original image and (W (m, n)) is the distorted image that contains (A*B) pixels.

Root Mean Square Error (MSE)

RMSE is measured as the square root of MSE. Mathematically it is defined as follows:

$$RMSE = \sqrt{MSE} \quad (4)$$

Peak Signal To Noise Ratio (PSNR)

The measurement of quality of reconstruction in compressed image is called PSNR. As PSNR value approaches to infinity. Higher image quality is obtained by getting the higher PSNR value. Mathematically it is defined as follows:

$$\text{PSNR} = 20 \log_{10}\left(\frac{(MAX)^2}{MSE}\right) \quad (5)$$

where MAX represents maximum fluctuation in image data type. Higher PSNR value, a lower RMSE and MSE value represents better image quality.

4.3 Segmentation

Image segmentation means to change the vision of image to make it more clear, meaningful and easy to analyze. Various techniques are being used for image segmentation. Thresholding method, edge detection based technique, region based technique, clustering based technique, artificial neural network based technique etc. In this research work, clustering based technique is used to perform segmentation. Clustering means segmenting the image into clusters having similar characteristics. Basically there are two types of clustering - Hard Clustering and Soft Clustering. Hard Clustering is the clustering where each data point is either completely the member of cluster or not. Soft Clustering where data point is a member of more than one cluster. Soft Clustering is used as this clustering is of natural type. Under soft clustering, we used fuzzy c-means algorithm as it uses partial membership therefore useful for real problems. This technique is also suitable in the field of medical imaging. The classification of clustering can be shown in figure 4.4.

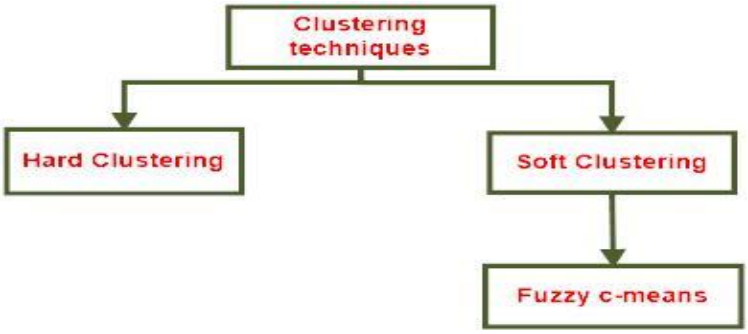


Figure 4.4: Clustering Techniques

4.3.1 Fuzzy c-means clustering

Fuzzy c means clustering is type of soft clustering and it is that type of clustering where one data belongs to two or more clusters**.** It has following function

$$K_m = \sum_{x=1}^{M} \sum_{y=1}^{N} {}_{xy}^{m} \left\| i_x - c_y \right\|^2 , 1<=m< \quad (6)$$

Where m is real number greater than one and $_{xy}$ represents degree of membership of i_x in the cluster y. i_x represents xth of d-dimensional measured data, c_y is the d-dimension center of the cluster. Figure 4.5 explains the workflow of the algorithm of fuzzy c means clustering.

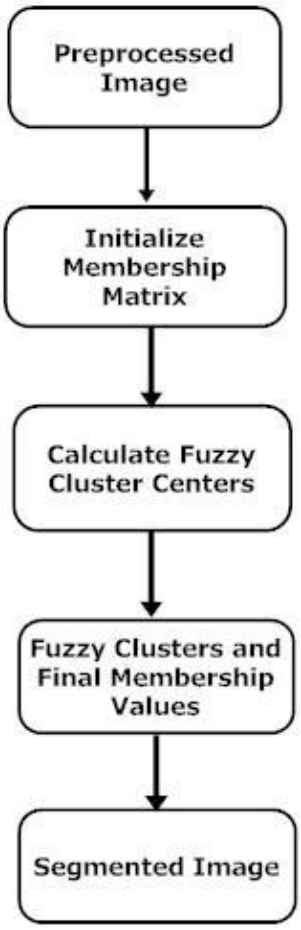


Figure 4.5: Flowchart of Fuzzy c means algorithm

4.4 Feature Extraction

Features have a very important role in image processing. Before extracting features, preprocessing and segmentation are applied on the image. After that for classifying and recognizing images, feature extraction techniques are applied. Feature extraction is the process of collecting higher information from an image which improves the accuracy of the system. In this research, Grey Level Co- Occurrence Matrix (GLCM) features are extracted. GLCM features are calculated by determining that how frequent pixel with value m happen horizontal to pixel with intensity value n. GLCM technique has following two steps:

1. GLCM is calculated.
2. Based on GLCM textures features are calculated.

GLCM is basically second order statistics therefore GLCM collected information regarding pixel of pairs. For creating GLCM, graycomatrix function is used. The figure 4.6 illustrates how these GLCM values are calculated. The output of element (1, 1) has value 1 because there is only one instance corresponding to input image. Similarly (1, 2) has value 2 corresponding to input image as there are two instances and so on.

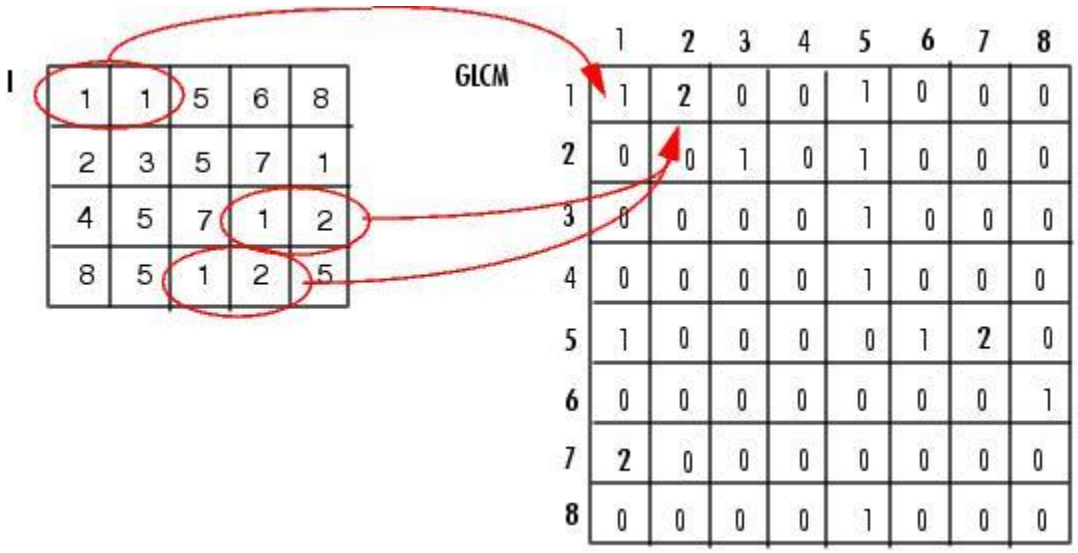


Figure 4.6: Construction of Gray Level Co-occurrence Matrix

GLCM is a process to determine how many different combinations they obtain of gray levels. Co-occurrence matrix works in four different orientations which are shown in figure 4.7. In this figure to calculate texture features every image block has been used and matrix which we get is made up at distance d=1 and at different angles i.e. 0 , 45 , 90 , 135 .The integer value that shows the distance between pixel and its neighboring is shown in figure 4.7 and same method is used for other image blocks.

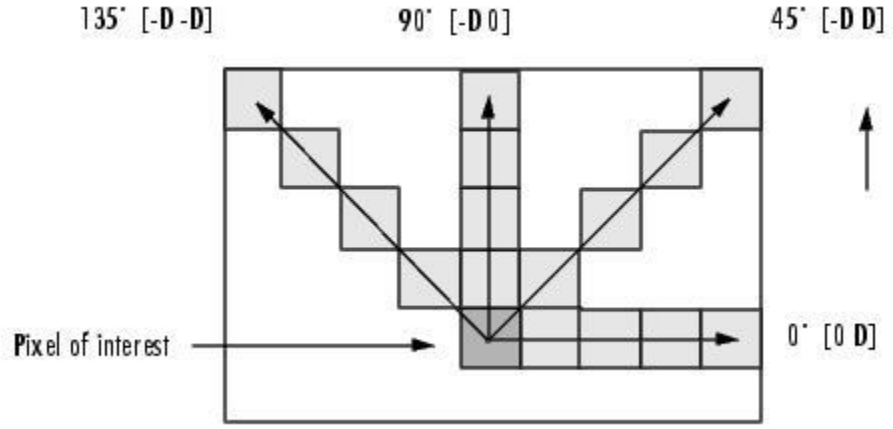


Figure 4.7: Different orientations of pixels

The features that we extracted in our research work are some first order statistics features and some are second order statistics features. Features that are based on texture of statistics are very helpful in medical images as they give non redundant and distinguishable features as comparison to other feature extraction methods.

1. First Order Statistics – These are the histogram based features which includes mean, variance, skewness, kurtosis, entropy. First order histogram P (I) can be defined based on below formula.

   P (I) = No. of pixels with grey level I / Total no. of pixels in the region (7)

   Following are the first order statistical features.

Mean (M) - Mean is calculated using sum of all values of pixel and divide them by total no of grey level values.

$$\frac{1^{x\,1\,y\,1}}{x^{*}\,y}\,R(m,n)(8)_{m0n0}$$

Variance (var) – It is used to measure the roughness from the image.

$$\frac{1}{x * y} \sum_{m=0}^{x-1} \sum_{n=0}^{y-1} (R(m, n))^2 \quad (9)$$

Skewness(S_k) - It is used to measure the lack of symmetry from the image over mean.

$$\frac{1(R(m, n))^3}{x * y \quad (\sqrt{\text{var}})^3} \quad (10)$$

Kurtosis(S_k) - It is used to describe the uniformity distribution of pixels of image.

$$\frac{1(R(m, n))^3}{x * y \quad (\sqrt{\text{var}})^4} \quad (11)$$

Entropy (E) - It is used to measure the randomness of the image and if R_{mn} is 0 then entropy is 0 and if R_{mn}'s are equal then entropy is maximum.

$$-\ln(R_{mn}) R_{mn} \quad (12)$$

2. Second Order Statistics - The features that are obtained from first order statistics gives only information about grey level distribution of an image only. Second order statistics gives information about relative positions of grey level within image. These features include Contrast, Correlation, Energy, and Homogeneity. Following are the first order statistical features.

   Contrast (c_{on}) – It is used to measure the intensity between pixel and neighbor of the image.

$$\sum_{m,n=0}^{P-1} R(m, n) * (m - n)_2 \quad (13)$$

   Correlation (c^{or}) - It is used to measure the correlation between pixel and neighbor of the image.

$$\sum_{m,n=0}^{P-1} K_{mn} \frac{(m)(n)}{2} \tag{14}$$

Energy (En) – It is used to measure the similarity of an image.

$$\sum_{m,n=0}^{P-1} (R_{mn})^2 \tag{15}$$

Homogeneity (H) – It describes the closeness of the image.

$$\sum_{m,n=0}^{P-1} \frac{R_{mn}}{1 (m n)^2} \tag{16}$$

where,

x * y = dimension and total no of pixels

m = Represents no of rows in matrix

n = Represents no of columns in matrix

R_{mn} = Element m, n of normalized symmetric geometry and

P = No of grey levels in image

GLCM mean

4.5 Classification

The GLCM features extracted are collected in a table. The classification of the images for the head and neck cancer is done by the classifiers such as Deep learning in h2o and decision tree. The brief discussion about the two classifiers is given in the subsequent sections.

4.5.1 Deep learning in h2o

In recent machine learning competitions, deep learning has been dominating for predictions with high accuracy. Deep learning belongs to the family of traditional machine learning methods for learning from data representations. Deep Learning [1] is a subpart of machine learning concerning with algorithms which are inspired by the brain

structuring and functioning called artificial neural networks. The major difference between deep learning and other machine learning algorithms is that deep learning can handle large amount of data without deteriorating the performance unlike other algorithms. In the proposed work, deep learning with H2O has been used. H2O [2] is fast scalable open source deep learning and machine learning for use in smarter applications. The enterprises like Cisco, Nielson, paypal etc. can use their whole data without sampling with H2O and can have better and faster predictions. The basic architecture of deep learning can be shown in Figure 4.8 where there are multiple hidden layers in contrast to artificial neural network where there is only single hidden layer. Therefore it can handle large amount of data along with individual features of data.

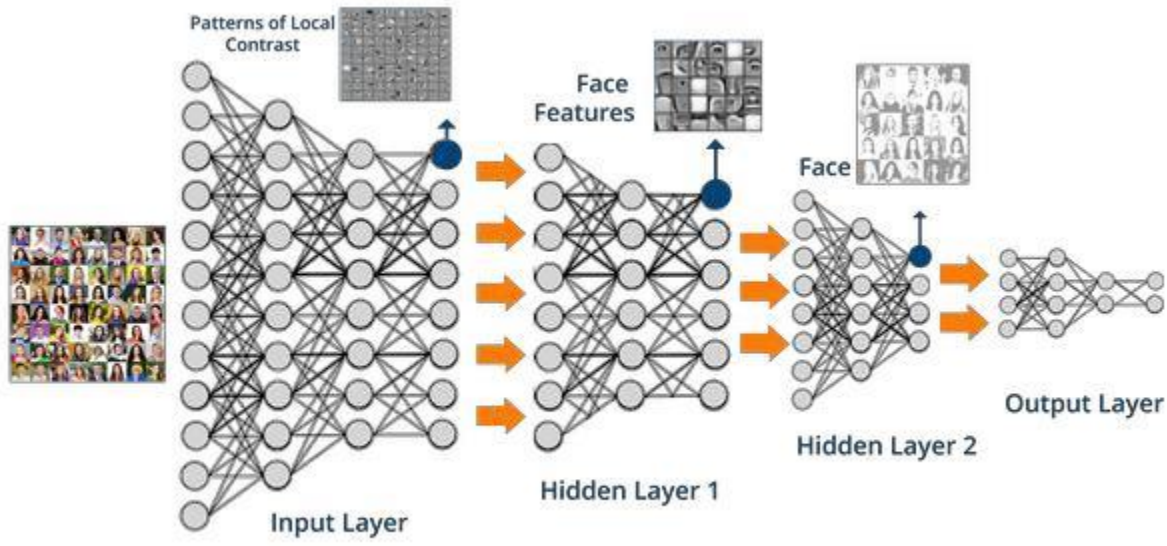


Figure 4.8: Deep Learning neural network

For deep learning there are several frameworks but here we used deep learning framework in h2o. This framework is based on multilayer feedforward neural network as shown in figure 4.9. This framework is based on the study of supervised training protocol and is applied to both regression and classification models. In this algorithm h2o.init() function is used to identify IP port, no. of threads to be used and we can also set the amount of RAM used in h2o. Out of 26019 data, 80% is used for training and 20% for testing and then train the model with appropriate parameters using h2o.deeplearning()

function. Deep learning with h2o algorithms performance is evaluated in terms of Specificity, Sensitivity and Accuracy.

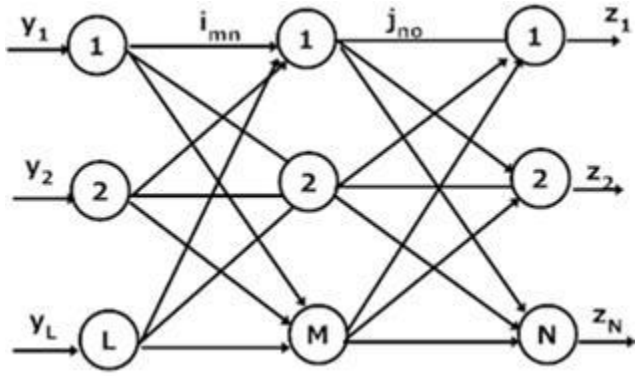


Figure 4.9: Multilayer Feed Forward Neural Network

$$_{n} \quad i \; i \quad y \quad _{0n \; m \, n} \quad _{m} \quad _{n} \tag{17}$$

$$z_{N} \quad j \; x \quad _{0o} \quad _{no} \quad _{no} \tag{18}$$

$$x_n \quad \frac{1}{1\exp(\quad _n)} \tag{19}$$

4.5.2 Decision Tree

Decision Tree comes under supervised learning algorithm and used for classification type of problem. This technique is useful where we have to segregate elements based on given input that is highly significant. Basic diagram of decision tree which shows important terminologies is shown in figure 4.10. Decision tree has basically of two types:

1. Categorical Variable Decision Tree – A tree whose target value is categorical is called categorical variable decision tree. Categorical means if value is in Yes or No.
2. Continuous Variable Decision Tree - A tree whose target value is continuous is called continuous variable decision tree.

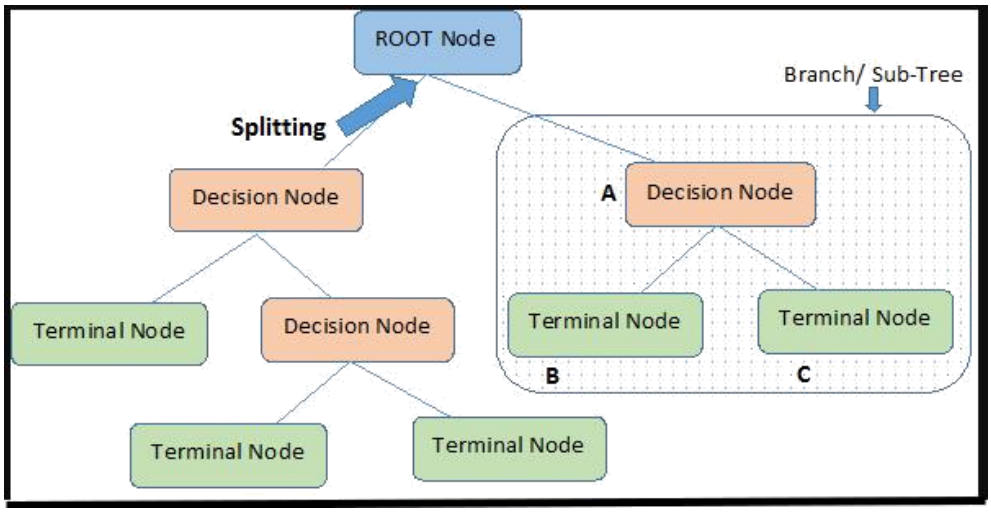


Figure 4.10: Basic Terminologies of decision tree

Here, out of 26019 data, 80% is used for training and 20% for testing and then train the model with appropriate parameters. Decision tree algorithm performance is evaluated in terms of Specificity, Sensitivity and Accuracy.

4.5.3 Performance Metrics for Classification Problems

The various performance metrics used for classification are confusion matrix, accuracy, precision, recall, specificity which can be explained as:

1. **Confusion Matrix**

   Confusion Matrix calculates the correctness and accuracy of the model. It comes under classification problem and used in either binary classification or multi-classification problem means output is either of two types or more types. For example – If a person having cancer or not is predicted by 1 or 0 value. If person suffers from cancer then value is 1 otherwise 0. Confusion Matrix table has two classes "Actual and Predicted". Various terms are associated with the Confusion Matrix and by using these terms we can calculate accuracy. Confusion Matrix is shown in table 4.1

Table 4.1: Confusion Matrix

Actual values	**Predicted values**		**Row Total**
	0	1	
0	TN	FP	TN+FP
1	FN	TP	FN+TP
Total	TN+FN	FP+TP	TP+FN+FP+TN

Terms Associated with Confusion Matrix

True Positive (TP): It is the case where actual and predicted class both are True or 1.

True Negative (TN): It is the case where actual and predicted class both are False or 0.

False Positive (FP): It is the case where actual class is False or 0 and predicted class is True or 1.

False Negative (FN): It is the case where actual class is True or 1 and predicted class is False or 0.

2. Accuracy

Accuracy is the sum of all the correct predictions divided by sum of all the predictions. Accuracy representation can be shown in figure 4.11. The accuracy can be calculated as:

$$\text{Accuracy} = \frac{TP \quad TN}{TP \quad FP \quad FN \quad TN} \tag{20}$$

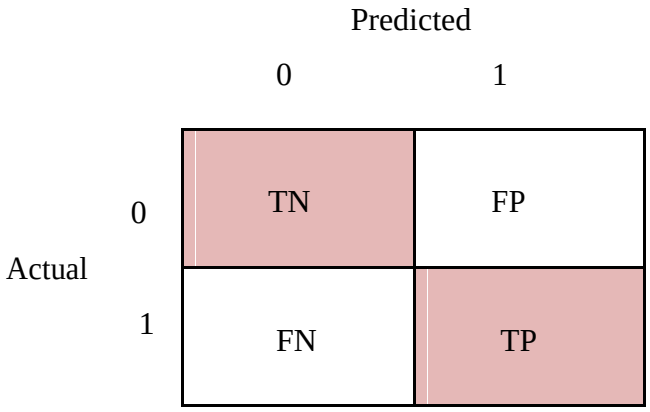


Figure 4.11: Accuracy representation

3. Precision

Precision measures what proportion of values that are predicted as true are actually true as shown in figure 4.12. Precision can be represented as:

$$\text{Precision} = \frac{TP}{TP \quad FP} \tag{21}$$

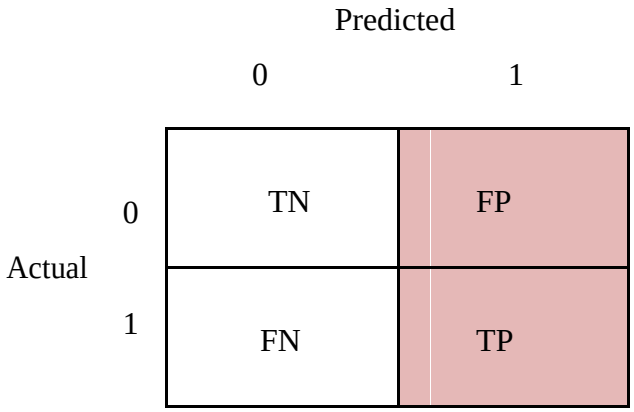


Figure 4.12: Precision Representation

4. Recall or Sensitivity

It measures what proportion of values that are actually true, are predicted as true as shown in figure 4.13. Recall can be represented as:

$$\text{Recall} = \frac{TP}{TP \quad FN} \tag{22}$$

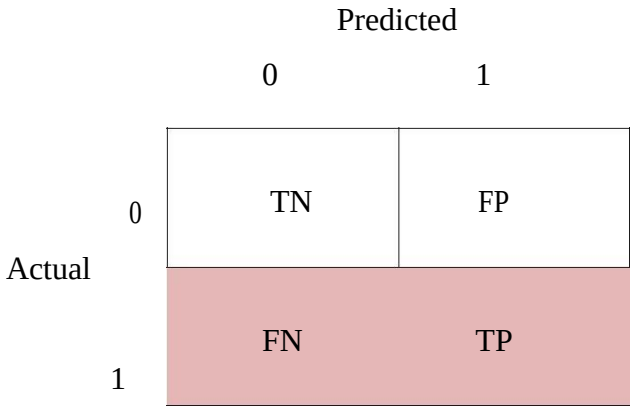


Figure 4.13: Recall Representation

5. **Specificity**

   It measures what proportion of values that were actually false, were predicted as false. It can be shown in figure 4.14. Specificity can be represented as:

$$\text{Specificity} = \frac{TN}{TN \quad FP} \tag{23}$$

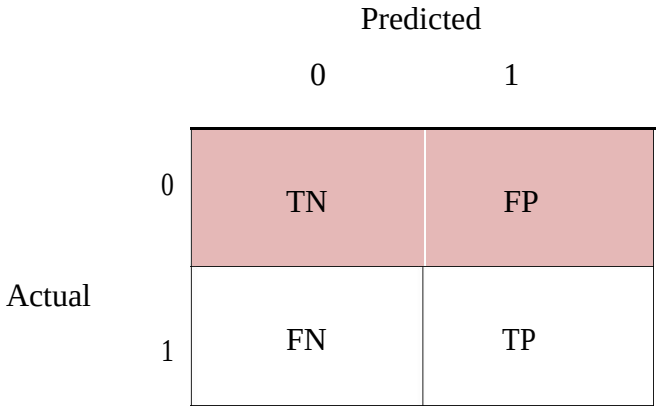


Figure 4.14: Specificity representation

4.6 Ensemble of models

Ensembling means combine the overall predictions obtained from the various machine learning algorithms to make stronger prediction. For ensembling there are various methods like averaging, majority voting, and weighed average but in this research work, averaging method is used and all the evaluation parameters are calculated for this. Ensembling of models is shown in figure 4.15.

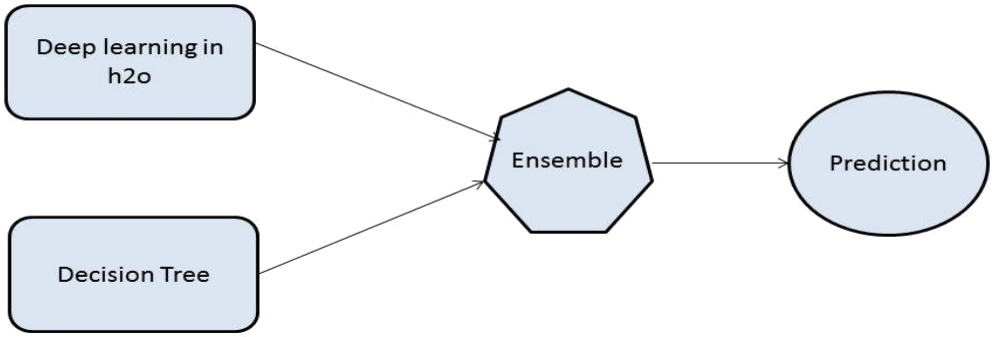


Figure 4.15: Ensembling of models

4.6.1 Types of Ensembling

1. **Averaging**

   Averaging method is best suitable for large problems and best method to reduce overfit. It is done by taking predicting probabilities. For example, as shown in table 4.2 the average prediction can be calculated as 35.

Table 4.2: Averaging method

Model1	Model2	Average Prediction
45	30	35

2. **Majority vote**

   It takes the prediction that has maximum vote in classification problem as shown in table 4.3.

Table 4.3: Majority voting

Model1	Model2	Model3	Voting Prediction
1	0	1	1

3. Weighted Average

Here weights and predictions of multiple models are taken and then average them for giving importance to model output as shown in table 4.4.

Table 4.4: Weighted Average

	Model1	Model2	Model3	Weighted Prediction
Weight	0.4	0.3	0.3	
Prediction	45	40	60	48

CHAPTER 5

RESULT ANALYSIS

In our research work, we used MATLAB R2017a for image processing and feature extraction and RStudio 3.4.4 for classification and we have taken the results on system having processor Intel Core i3-2348M CPU @2.30GHZ, 4GB RAM and 64-bit Operating System. There are 26019 images collected from Cancer Imaging Archive (TCIA) and images are in dicom format. Dicom format images are mainly for the medical related data.

After preprocessing results are obtained using various filtering techniques like weiner filter, average filter, median filter and Guassian filter and comparison parameters like MSE, PSNR and RMSE are computed. The output images of median filter and weiner filter is shown in figure 5.1 and average and Guassian filter o/p images are shown in figure 5.2. Table 5.1 compares the results from all the four filters. The figure 5.3 depicts the comparison of the four filters graphically.

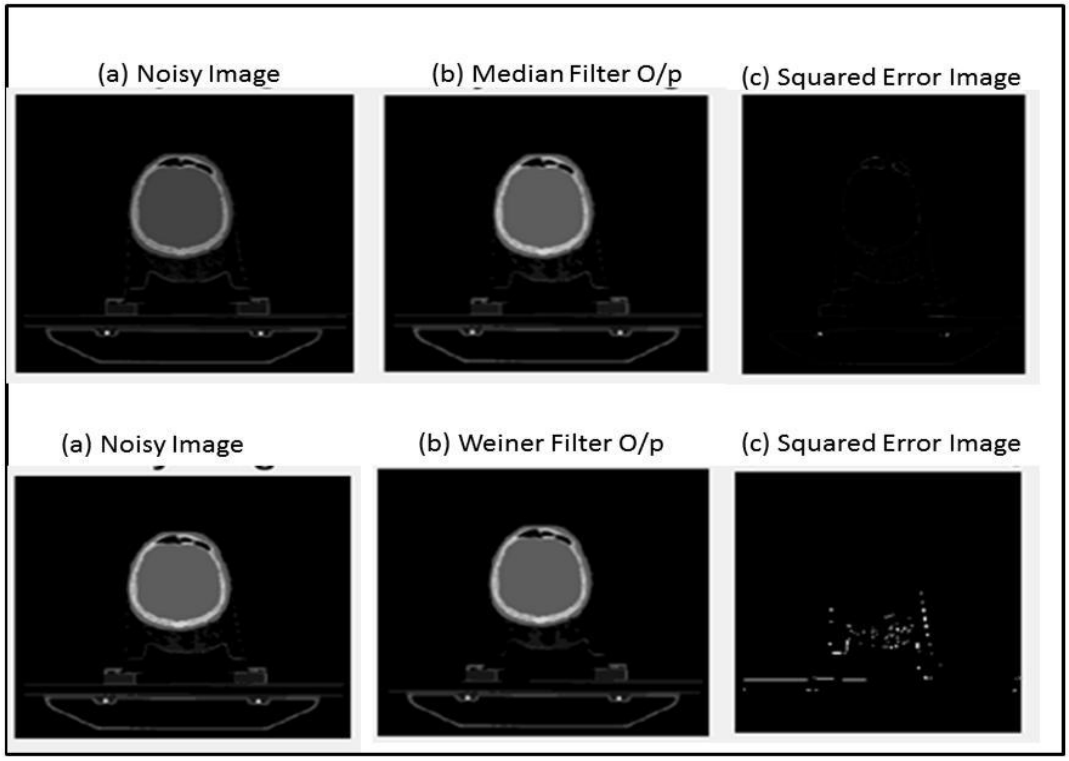


Figure 5.1: Image 1 – Results of median and weiner filter o/p

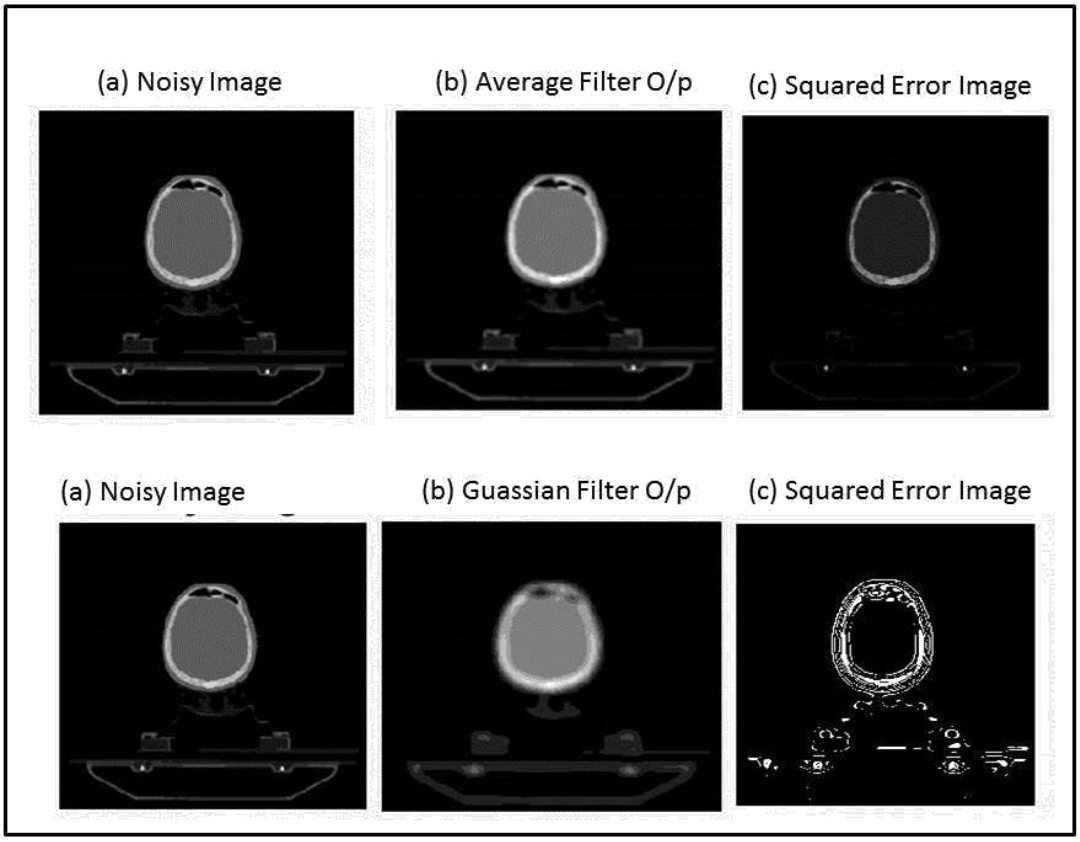


Figure 5.2: Image 1 - Results of Average filter and Guassian Filter Output

Table 5.1: Comparison Parameters for image 1

Comparison Parameters	PSNR	MSE	RMSE
Median Filter	76.2012	0.0016	0.0396
Weiner Filter	96.3296	1.52596e-05	0.0039
Average Filter	45.1909	1.9833	1.4083
Guassian Filter	62.4167	0.0376	0.1938

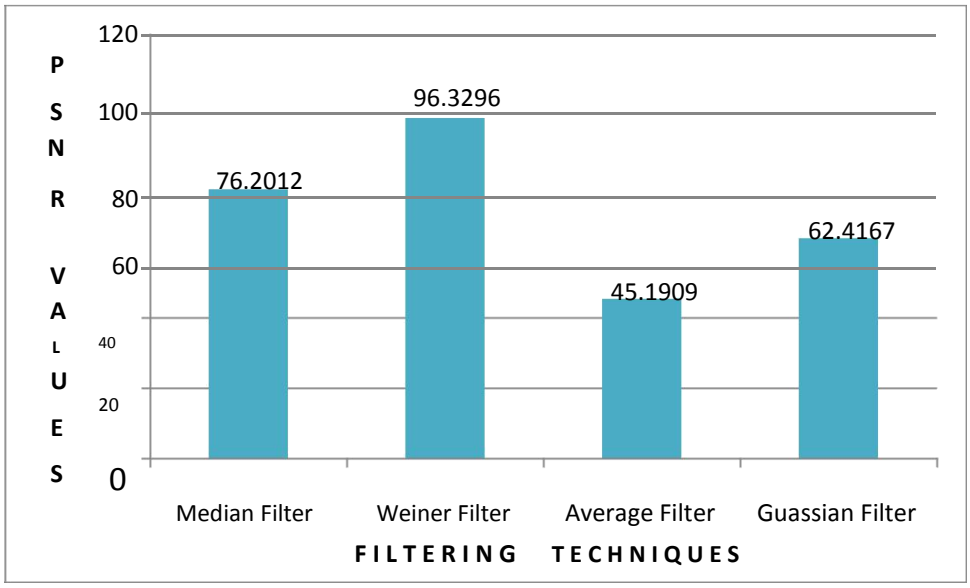


Figure 5.3: The comparison of various filtering Techniques

From the table 5.1, it is clear that weiner filter is the best technique for noise filtering as this technique has minimum MSE and RMSE value and maximum PSNR value. Therefore after comparison weiner filter technique is used for preprocessing and further on preprocessed images segmentation is done on the images using fuzzy c means clustering method. The original image, preprocessed image and segmented image obtained is shown in figure 5.4.

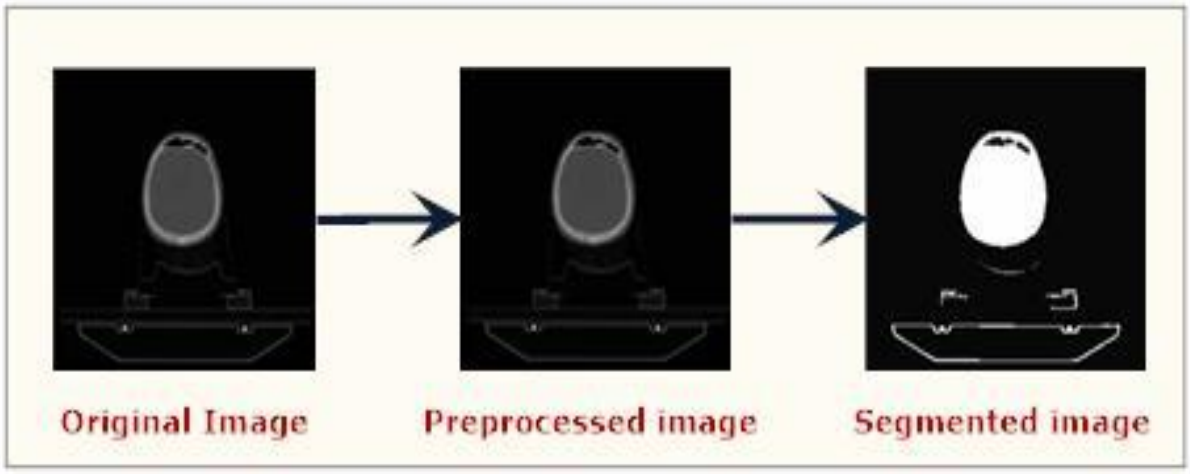


Figure 5.4: Original, Preprocessed and Segmented Images

After segmentation, feature extraction is done. Table 5.2 and Table 5.3 shows the extracted features of few images.

Table 5.2: First order Statistical Feature of few images

Images	Mean	Variance	Skewness	Kurtosis	Entropy
1	8.22781	5.530861	-2.73054	8.46705	0.35502
2	5.98333	5.347266	-1.64346	3.84055	0.61062
3	8.06441	6.540969	-2.37997	6.67503	0.408055
4	4.92141	8.281385	-0.61724	1.40349	0.756003
5	8.29266	5.089839	-2.89372	9.39146	0.345419
6	7.13629	3.452225	-2.12538	8.13951	0.871992

Table 5.3: Second order statistical feature of few images

Images	Contrast	Correlation	Energy	Homogeneity
1	0.246728	0.971207	0.81953	0.995331
2	0.283184	0.973118	0.65476	0.986204
3	0.392742	0.961345	0.78417	0.992644
4	0.325312	0.980112	0.50371	0.981793
5	0.383309	0.951697	0.82962	0.992771
6	0.328339	0.938923	0.49483	0.985514

After feature extraction to classify the image as cancer or not, no of white pixels are calculated from the images and based on that we get cancer and non-cancer images. Figure 5.5 depicts that the frequency of cancer images (data class 1) is more than non-cancer images (data class 0) in our dataset.

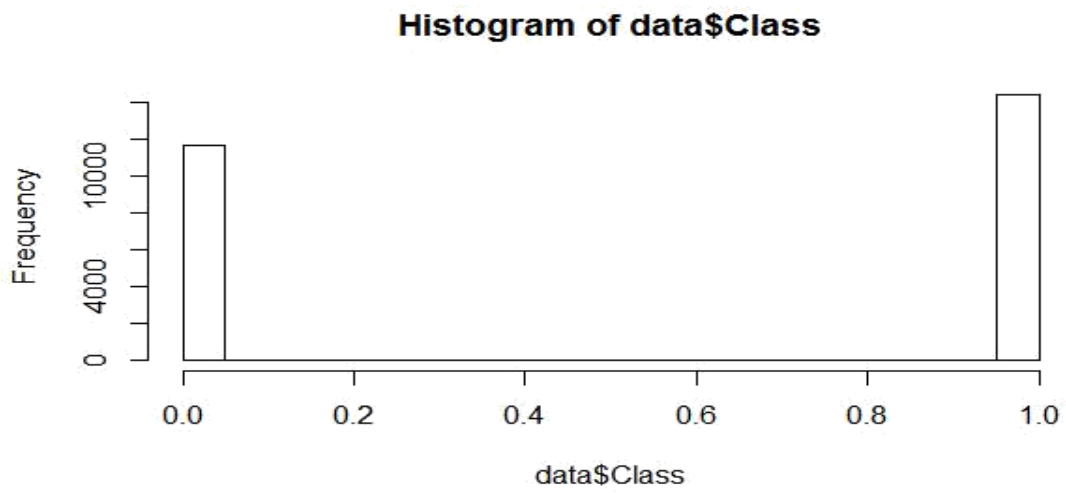


Figure 5.5: Histogram of Class

Now these extracted features are used to train the models which we used in our research. Following are the results that are obtained from models used in research work.

5.1 Deep learning in h2o

Here we used deep learning framework in h2o to train and test the features where 80% data is used to train the data and 20% is used to test the data. The confusion matrix of both test data and train data are shown in table 5.4 and table 5.5 and results obtained from experiments for train data have 98.8% accuracy, 98.7% specificity, 98.8% sensitivity and 98.9% precision as shown in table 5.6 and test data have 98.9% accuracy, 99.3% specificity, 98.6% sensitivity and 99.5% precision as shown in table 5.7 . We get ROC curve graph shown below in figure 5.6 which is used to describe specificity and sensitivity tradeoff for binary classifier.

Table 5.4: Confusion Matrix for Train Data

n=20917	0(Predicted)	1(Predicted)
0(Actual)	9246	119
1(Actual)	128	11424

Table 5.5: Confusion Matrix for Test data

n=5102	0(Predicted)	1(Predicted)
0(Actual)	2264	14
1(Actual)	38	2786

Table 5.6: Model Performance Parameters for Train Data

Accuracy	98.8%
Sensitivity(Recall)	98.8%
Specificity	98.7%
Precision	98.9%

Table 5.7: Model Performance Parameters of Test data

Accuracy	98.9%
Sensitivity(Recall)	98.6%
Specificity	99.3%
Precision	99.5%

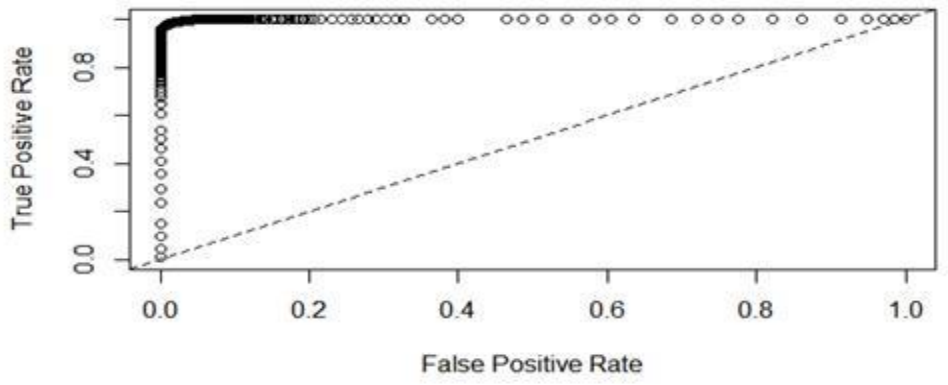


Figure 5.6: ROC curve

5.2 Decision Tree

Here we used decision tree to train and test the features where 80% data is used to train the data and 20% is used to test the data. Decision tree is shown in figure 5.7. From this figure we can observe that variance is the most important feature among all features which helps to classify whether there is cancer or not and if variance value is less than equal to 6.909 then move to mean feature and further if value is less than equal to 6.006 then reaches at leaf node which shows that 99% suffer from cancer and this same procedure will continue in whole tree. The confusion matrix of both test data and train data are shown in table 5.8 and table 5.9 and results obtained from experiments for train data have 98.4% accuracy, 98.9% specificity, 98.8% sensitivity and 98.8% precision as shown in table 5.10 and test data have 98.6% accuracy, 99.9% specificity, 98.9% sensitivity and 99.9% precision as shown in Table 5.11.

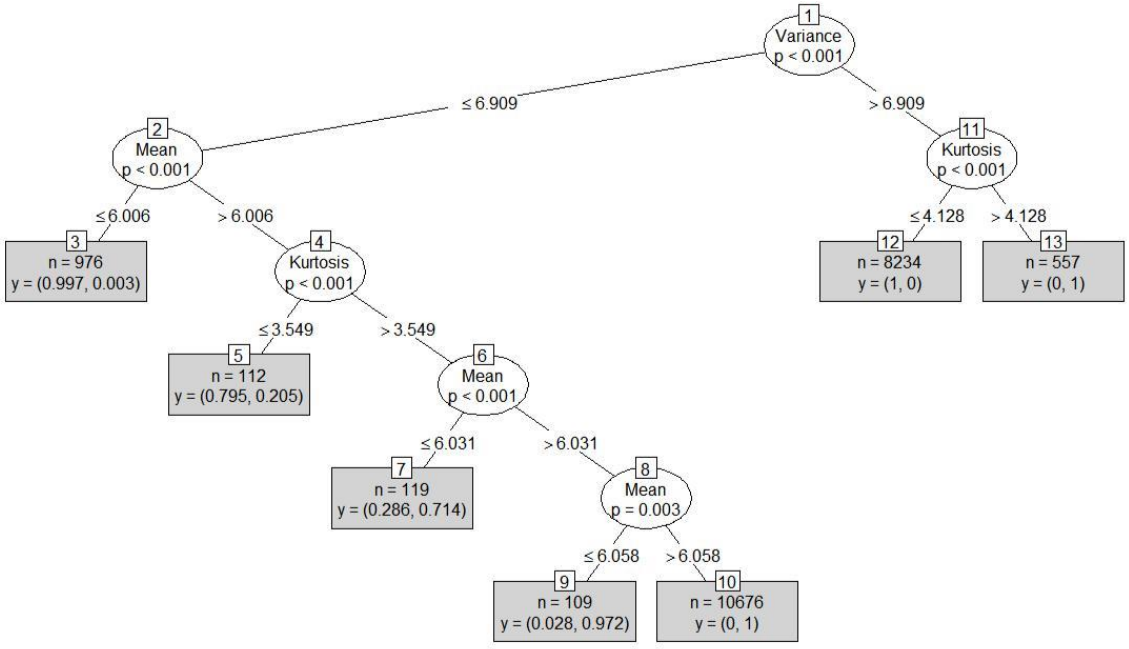


Figure 5.7: Decision Tree for training data

Table 5.8 Confusion Matrix for Train Data

n=20783	0(Predicted)	1(Predicted)
0(Actual)	9061	181
1(Actual)	361	11180

Table 5.9: Confusion Matrix for Test data

n=5236	0(Predicted)	1(Predicted)
0(Actual)	2249	24
1(Actual)	92	2871

Table 5.10: Model Performance Parameters of Train Data

Accuracy	97.39%
Sensitivity(Recall)	96.8%
Specificity	98.04%
Precision	98.4%

Table 5.11: Model Performance Parameters of Test data

Accuracy	97.78%
Sensitivity(Recall)	96.8%
Specificity	98.9%
Precision	99.17%

5.3 Ensembled Results

Results after ensembling of deep learning in h2o and decision tree is shown in table 5.12

Table 5.12 Ensembled Model

Model Name	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
Deep learning in h20 and decision tree	99.41	99.1%	99.7%	99.9%

5.4 Comparative Analysis

The results of research work is compared with Support Vector Machine (SVM), K-NN Classifier and Logistic Model (LR) on the basis of accuracy and from our results it is seen that proposed method gives better results than all three models. Table 5.13 and Figure 5.8 shows comparison of different classifiers.

Table 5.13: Comparison of different classifiers based on accuracy

Classifiers	Accuracy	Specificity	Sensitivity
Ensembled Model	99.41%	99.7%	99.1%
Deep learning in h2o	98.9%	99.3%	98.6%
Decision Tree	97.78%	98.9%	96.8%
SVM	99.3%	98.85%	66.58%
kNN	89.4%	86.9%	91.9%
LR	81.8%	82.2%	81.4%

From the table 5.13 it can be observed that the proposed ensemble model outperforms the state-of-the-art techniques

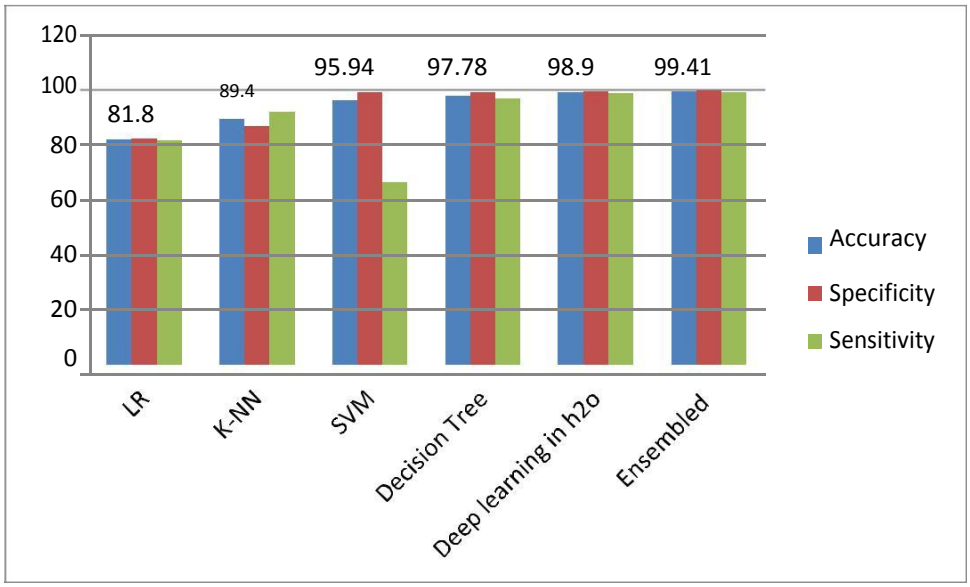


Figure 5.8: Graphical representations of different classifiers

The Figure 5.8 show the graphical comparison of the exiting methods with proposed methods for the detection of head and neck cancer by comparing parameters accuracy, sensitivity and specificity. Existing approaches used one single model which did not give stronger results. Proposed approach provides the Ensembled model consisting of two models as base learners which give stronger overall prediction.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Head and neck cancer is common nowadays. Initially very less people suffered from this cancer. This cancer has about average of 5 years life expectancy. Cancer is treated by radiation therapy or surgery but these radiations can affect the health of patient and reduces the life of the patient. We cannot take it lightly therefore we have to solve this problem by using new technologies like Artificial Intelligence and Machine Learning. By taking only Computed Tomography (CT) scan, Magnetic Resource Imaging (MRI) or Positron Emission Tomography (PET) scan images where we can predict cancer using machine learning algorithms. This research work predicts whether the people has cancer or not with very high accuracy by extracting textural features as textural features are useful for medical data.

In this research work, using CT scan images, we detect head and neck cancer by taking 26019 images. The weiner filter has been used to remove noise from images and is better than other methods. Then after preprocessing to get more clarity in image or to deeply understand the meaning of image, segmentation using fuzzy c means clustering method is done. Further feature extraction is done using texture features and histogram based features. Then, target is calculated from extracted features based on white pixels and system is trained using models deep learning in h2o and decision tree where 80% data is used for training and 20% data is used for testing. From these two models we achieved accuracy of 98.8% and 99.4%. By this we investigated that this technique is better than manual detection that is performed clinically. Then ensembling is done on these two models and achieve overall higher accuracy of 99.41%.

The major contributions of thesis are:

1. The preprocessing technique has been used to enhance the images for better results.

2. CT scan images are used to train the machine learning model to achieve better accuracy.
3. Ensembling of deep learning in h2o and decision tree has been done to achieve accuracy of 99.41%.

6.2 Future Work

In future we can use any feature selection technique by which the results can be improved. The proposed methodology can also be applied to detect lung cancer, breast cancer, brain tumour etc. PET scan images dataset can be used instead of CT scan images.

REFERENCES

[1] Radiation Therapy for Head and Neck Cancers, available at http://www.icradonc.com/treatment/disease/head_neck.htm, accessed on 6th June, 2018.

[2] D.W. Tshering Vogel and H.C. Thoeny, "Cross-sectional imaging in cancers of the head and neck: how we review and report", Cancer Imaging, Vol. 16, no. 20, pp. 1-15, 2016

[3] Salivary Glands & Thyroid Cancer available at http://www.beaconhospital.com.my/salivary-glands-thyroid-cancer, accessed on 29th

[4] Head And Neck Cancer, available at https://www.slideshare.net/doctorbobm/head-and-neck-cancer, accessed on 30th May, 2018.

[5] Hermans, R. (2006). Head and Neck cancer Imaging.

[6] M. Mete, X. Xu and C. Fan, "A Machine Learning Approach for Identification of Head and Neck Squamous Cell Carcinoma," *In the Proceedings of IEEE Bioinformatics and Biomedicine*, vol.4, no.1, pp.29-34, 2007.

[7] M. Kulkarni, "Head and Neck Cancer Burden in India," *In the Proceedings of International Journal of Head and Neck Surgery*, vol. 4, no.1, pp.29-35, 2013.

[8] A. Argiris, M. Karamouzis, D. Raben, and R.Ferris, "Head and neck cancer," *The Lancet*, vol. 371, no. 9625, pp.1695–1709, 2008**.**

[9] N. Vigneswara and M. D. Williams, "*Epidemiologic Trends in Head and Neck Cancer and Aids in Diagnosis*," in Oral Maxillofacial Surg Clin N Am, vol.26, no.2, pp.123-141, 2014.

[10] Head and neck cancer, available at https://www.macmillan.org.uk/information-and-support/head-and-neck-cancers/understanding-cancer, accessed on 6th June, 2018.

[11] National Cancer Registry Programme (ICMR) 2009. Consolidated Report of HBCR: 2004-2006, Bangalore, India.

[12] M. Nagu and N. Shankar, "Image De-Noising By Using Median Filter and Weiner Filter," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 9, pp. 5641-5649, 2014.

[13] S.Perumal and T.Velmurugan, "Preprocessing by Contrast Enhancement Techniques for Medical Images," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 18, pp. 3681-3688, 2018.

[14] Y. Labeeb and Dr. M. Morsy, "Preprocessing Technique for Enhancing the DICOM Kidney Images," *International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 08, pp. 836-841, 2015.

[15] T. Chhabra, G. Dua, T. Malhotra, "Comparative Analysis of Methods to De-noise CT Scan Images," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 7, pp. 3363-3369, 2013.

[16] D. Kaur, "Various Image Segmentation Techniques: A Review," *International Journal of Computer Science and Mobile Computing*, vol.3 no.5, pp. 809-814, 2014.

[17] A. Norouzi, "Medical Image Segmentation Methods, Algorithms, and Application," *IETE Technical Review*, vol. 31, no. 3, pp. 199-213, 2014.

[18] W. Ahmad and M. Fauzi, "Comparison of Different Feature Extraction Techniques in Content-Based Image Retrieval for CT Brain Images", *IEEE 10th Workshop on Multimedia Signal Processing*, pp. 503-508, 2008

[19] S. Singh , Y. Singh and R. Vijay, "An Evaluation of Features Extraction from Lung CT Images for the Classification Stage of Malignancy", in *IOSR Journal of Computer Engineering*, pp. 78-83, 2016.

[20] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," *in Competition and cooperation in neural nets. Springer*, pp. 267–285, 1982

[21] L. Yann, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE Vol. 86, no. 11, pp. 2278-2324, 1998.

[22] S. Behnke, "Hierarchical neural networks for image interpretation", in Lecture Notes in Computer Science, vol. 2766, 2003.

[23] P. Y. Simard, D. Steinkraus, J. C. Platt , "Best practices for convolutional neural networks applied to visual document analysis." *in ICDAR*, vol. 3, pp. 958–962, 2003.

[24] R. Fakoor, F. Ladhak, A. Nazi and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," *in Proceedings of the International Conference on Machine Learning*, vol. 28, 2013.

[25] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[26] A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," *in International Conference on Medical Image Computing and Computer-Assisted Intervention*. *Springer*, pp. 403–410, 2013

[27] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[28] D. C. Cires, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," *in International Conference on Medical Image Computing and Computer-assisted Intervention. Springer*, pp. 411–418, 2013.

[29] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. Gonz´alez, and A. Madabhushi, "Accurate and reproducible invasive

breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent,'' *Scientific reports*, vol. 7, p. 46450, 2017.

[30] Y. K. Tsehay, N. S. Lay, H. R. Roth, X. Wang, J. T. Kwak, B. I. Turkbey, P. A. Pinto, B. J. Wood, and R. M. Summers, "Convolutional neural network based deep-learning architecture for prostate cancer detection on multi-parametric magnetic resonance images," *in Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, p. 1013405 , 2017

[31] G. Sujatha , Dr. K. Usha Rani, "Evaluation of Decision Tree Classifiers on Tumor Datasets", *International Journal of Emerging Trends & Technology in Computer Science*, vol. 2, no. 4, pp. 418-423, 2013.

[32] A. Elsayad, "Diagnosis of Breast Cancer using Decision Tree Models and SVM", *International Journal of Computer Applications,* vol. 83, no.5, pp.19-29, 2011

[33] P.Hamsagayathri and P.Sampath "Priority Based Decision Tree Classifier for Breast Cancer Detection", *2017 International Conference on Advanced Computing and Communication Systems.*

[34] H. Hijazi, M. Wu, A. Nath and C. Chan, "Ensemble Classification of Cancer Types and Biomarker Identification", *Drug Dev Res. 2012 Nov*, vol.73 , no 7 , pp. 414–419.

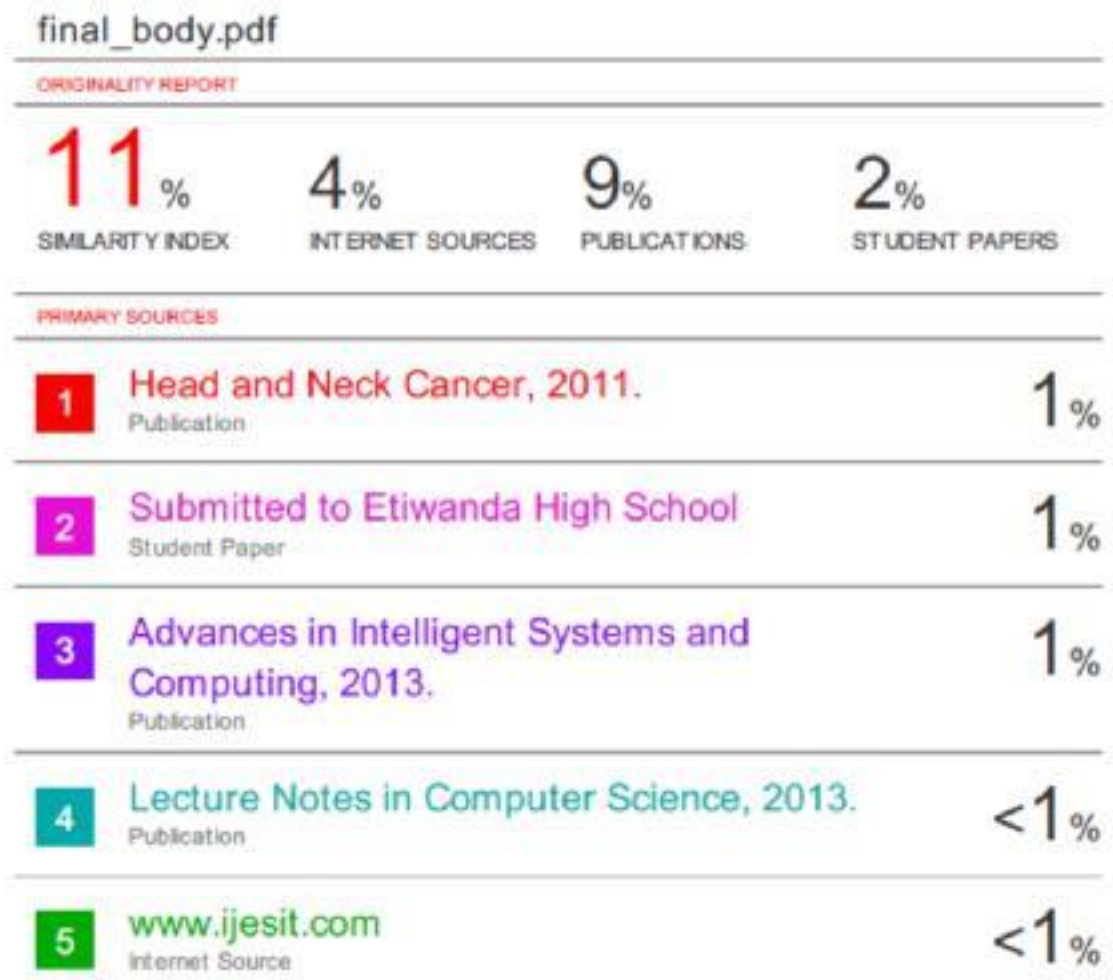
[35] T. G. Dietterich, "Ensemble methods in machine learning", *International workshop on multiple classifier systems,* vol. 33, pp. 1-15, 2000.

[36] A. Onan, "On the Performance of Ensemble Learning for Automated Diagnosis of Breast Cancer", *Artificial Intelligence Perspectives and Applications, Advances in Intelligent Systems and Computing*, pp. 119-129.

PUBLICATIONS

1. Pooja Gupta, Dr. Avleen Kaur Malhi, "Using deep learning to enhance head and neck cancer diagnosis and classification", *IEEE international conference on Systems, Computation, Automation and Networking(ICSCAN), 2018* [Accepted].

PLAGIARISM REPORT



final_body.pdf

ORIGINALITY REPORT

11%	4%	9%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Head and Neck Cancer, 2011. Publication	1%
2	Submitted to Etiwanda High School Student Paper	1%
3	Advances in Intelligent Systems and Computing, 2013. Publication	1%
4	Lecture Notes in Computer Science, 2013. Publication	<1%
5	www.ijesit.com Internet Source	<1%