Machine Learning Method for Pancreatic Cancer Detection using Naïve Bayes and Decision Tree Algorithm

R.Veeramani, Aryan Goswami, Harsh Aditya, Praveen Ranjan

Abstract: *Machine Learning is the study of algorithms and application of artificial intelligence. Artificial Intelligence is said to be the superset of machine learning. It aims to develop those programs which can learn and improves it upon increasing experience. It is designed to learn by itself. The aim is to detect pancreatic cancer using machine learning approach. Pancreas is responsible for secreting insulin which helps to control the blood glucose level in the human body. The paper aims to detect pancreatic cancer with the help of machine learning. The tumor is detected using image processing and is to be detected at the premature stage so that proper medication and treatment can be provided to increase the survival rate of the patient. The MRI image of pancreas obtained after MRI scan will be preprocessed and its noise is removed. The segmentation of MRI images will be performed using FCM algorithm. The tumor present in the image will be detected with the help of morphological process and multi clustering model. After Segmentation the image will be divided into various regions. With the help of the hybrid technique the primary and secondary regions are compressed and are used for telemedicine application. DWT is used for DE noising the image. GLCM features are extracted. The image then compared with the database images of pancreatic tumors and is classified as abnormal and normal with the help of BPN based classifier. The image is classified into abnormal and normal. The malignant image is considered as abnormal. The abnormal image is then segmented using SFCM and tumor part is clustered. After clustering the tumor part validation about the presence of pancreatic cancer is given.*

Keywords: *MRI (Magnetic Resonance Imaging), FCM (Fuzzy C means clustering), DWT (Discrete Wavelet transform), BPN (Back Propagation Network) Classifier, GLCM (Gray Level Co-occurrence Matrix) feature.*

I. INTRODUCTION

Machine Learning is an interpretive barbarized computing model which works on the principle that machines will learn from knowledge, establish the concerned patterns and takes call with minimum quantity of human interference. The most advantage of machine learning is that its accuracy will increase with increase in expertise. It learns from the previous computations to produce a trustable decision and result.

It is also used in disease identification and diagnosis, personalized treatment, drug discovery, epidemic outbreak and many more. In this paper we are going to predict pancreatic cancer using machine learning.

The disease which is the resultant of unrestricted growth of extra cells in body is termed as cancer. It happens when the normal control mechanism of body stops working. A tumor is formed when these extra cells form a mass of tissue. It can affect any organ of the body and one such organ is pancreas. Pancreas is an organ which is present across the back of the abdomen behind the stomach. It is about 6 inches long. Pancreatic cancer is formed when DNA of cells present in the pancreas gets changed. The functions which are to be performed by cells is ordered by DNA .The change in the DNA is termed as the mutation. The mutations order the cells to grow at an uncontrollable rate and continue to live after the death of normal cell. These cells accumulate together and form a tumor. Image processing technique is used to detect tumor. The MRI scan image of the pancreas is first taken and preprocessing of the image is done.

The noise which occurs due to motion blur, camera misfocus or due to any other problem is removed with the help of filters. The image is compared with the database images and a classification is done using BPN based classifier to classify the image as normal or malignant. The malignant image is again segmented using SFCM algorithm. The tumor region of the image is smoothed using morphological process

II. LITERATURE SURVEY

Cancer is a disease which involves abnormal cell growth and results in the destruction of body tissue. It is considered to be one of the major causes of death worldwide.

Pancreatic cancer is also the leading cause of death in men as well as women. In the reason for death worldwide it stands at the fourth position among all the disease. In the upcoming future it has been estimated by the scientists that it can be the second reason of death due to cancer in the United States. The disease is curable if the tumors of the same are detected at an early stage. The detection of the tumor with the help of the images is attempted through this paper.

Revised Manuscript Received on May 30, 2020.

***** Correspondence Author

R. Veeramani*, Assistant Professor, Dept. of IT, SRM IST, Chennai, Tamil Nadu, India.

Aryan Goswami, Student, Bachelor of Technology, Information Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Harsh Aditya Student, Bachelor of Technology, Information Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Praveen Ranjan, Student, Bachelor of Technology, Information Technology, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Numerous attempts have been already made to detect the tumor of the pancreatic cancer with the help of images .

In Paper [1] the detection of cancer with the help of tumor images was attempted.

Two processes which includes image processing technique as the first one and basic classifier as the other to detect the cancer causing tumor. After the completion of the image preprocessing technique minimum distance classifier algorithm was used for the tumor detection in the given image. The accuracy of the given process is considered to be of sixty percent. The scanned images of the tumor collected as datasets were given to the system as inputs. The Scanned image used to contain noise due to the various reason such as machine error, scanner error etc. Thus the important and foremost step was noise removal with the help of filters. After the removal of noise the classification of images is done using various parameters. The classification of organs from the image is done by the classifier. It helps to distinguish pancreas from other organs present in the image. Different organs are shaded with different colors in the image. The color representing the pancreas is separated for further process. Error is calculated using minimum standard deviation and maximum distance error. The tumor is detected in the image and the possibility of pancreatic cancer is confirmed.

In paper [2] the tumor present in the pancreas was detected using the Bacterial Foraging Algorithm. Features are extracted from the image with the help of equations. Features consist of contrast, homogeneity, energy and entropy are extracted. The features calculate number of repeated pairs, finding the texture of image with the help of randomness. The data collected in the form of these features is provided to the bacterial foraging algorithm to categorize whether the tumor is present in the pancreas or not. The paper mainly focuses on the usage of bacterial foraging algorithm. It consists of four stages which are chemo taxis, reproduction, elimination dispersal and swarming. The patients which have the symptoms of pancreatic cancer undergo pancreatic screening. It is done with the help of x-ray imaging or magnetic resonance image screening. The benign and malignant tumors in pancreas are detected with the help of Gaussian filter which removes noise. The values obtained are then normalized and used to identify the presence of tumor and possibility of pancreatic cancer is confirmed.

In paper [3] an attempt has been made to predict the cancer with the help of neural networks. Artificial Neural Network is said to be best in terms of sensitivity as compared to other available algorithms. In these neural network two extra layers with each one having twelve neurons. The best model is chosen after training on the dataset. A threshold value is set and if the model gets value above the threshold value then it yields ‘YES’ otherwise it will return the value as ‘NO’. Based on the datasets provided the risk analysis is done and it is also used to categorize which set of people are more risky and what are initial signs which is exhibited in most of the patients.

III. PROPOSED SYSTEM

The deaths due to pancreatic cancer are high in developed countries as compared to other countries. The reason behind it might be lack of diagnosis. The pancreatic cancer patients do not show symptoms at an early stage like other cancer. Over four hundred fifty thousand people have been affected from this deadly disease in the year 2018 all over the world. It has been estimated that over 400 thousand new cases will occur until 2050. The 5-year survival rate stands lowest in pancreatic cancer i.e. of 9% only. Therefore it becomes very important to detect cancer tumor at an early stage so that the proper diagnosis and treatment of the patient can be done and humanity can triumph over this deadly and fatal disease.

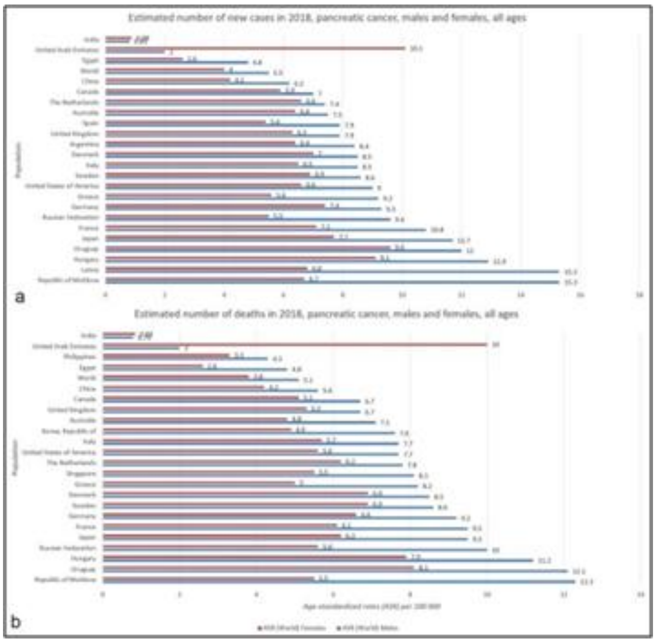

Fig. 1 Mortality Rate of Pancreatic Cancer 2018 worldwide

We proposed a method to detect pancreatic cancer in this paper.

- The Paper proposes to spot the tumor from MRI scanned medical images using multi clustering model and morphological process.
- The segmentation refers to the process of partitioning a digital image into multiple segments.
- The Pancreatic MRI is taken and its noises are removed using filters and then applied spatial Fuzzy C means Clustering algorithm for the segmentation of MRI images.
- The morphological process will be used to smooth the tumor region from the noisy background.
- The segmented primary and secondary regions are compressed with hybrid techniques for telemedicine application.

A. Discrete Wavelet Transform

DE noising is the main application of the DWT. It works on the principle of averaging several denoising signals. It is most suitable method for noise removal as well as signal compression with fewer coefficients. It enables a sparser representation.

Scaling and Translation are the initial steps in the working of the algorithm.

Scaling

2^j (j = 1, 2, 3, 4)

The Base scale in the discrete wavelet transform is fixed to two.

Different scales can be obtained when the root scale is raised to the integral numbers

2^j (m) (m= 1, 2, 3, 4)

Translation happens at whole number multiples drawn during this equation. It eliminates redundancy in coefficients.

It is often referred as dyadic scaling and shifting.

B. GLCM Features Extraction

It is the oldest statistical method of second order to analyze the texture of the image.

Following are the features of GLCM.

a. Angular Second Moment

It is also called as uniformity.

b. Entropy

c. Contrast

d. Homogeneity

It is also called as inverse difference moment.

e. Variance

C. Naïve Baye's And Decision Tree Classifier

Naïve Bayes is a classification technique based on Bayes theorem. The predictors perform their role independently. It consists of two parts which is Naïve and Bayes. It works on the principle that all the features are independent in their existence. They are not related to each other. In case the features are interdependent then also each one of them contributes independently to the probability.

Decision tree works on the principle of dividing root nodes into several nodes. Each node asks a true false question about one of the feature and the tree grows recursively on the basis of the questions asked.

D. FCM

Fuzzy means that regarding a style of pure mathematics and logic during which predicates could have degrees of pertinence instead of easy being true or false.

The central plan in fuzzy agglomeration is that the non-unique partitioning of the information in assortment of clusters. It's superset of mathematical logic.

It's referred to as soft agglomeration. it's a task of clustering set of knowledge points in such the simplest way that objects within the same group square measure a lot of like {each different one another} than to those in other teams. It's the common technique for applied math knowledge analysis and employed in several fields like pattern recognition, image analysis, knowledge compression, data retrieval and camera work.

IV. SYSTEM ARCHITECTURE

The system comprises of three phases which will detect presence of pancreatic tumor in the image.

The first step is the preprocessing phase in which the image is first preprocessed to suppress the abnormalities present in the image which may create some problem in getting the desired output. It also focuses on enhancing some of the features of the image which are crucial for further processes.

The discrete wavelet transform is used to remove the noise from the image.

Noise may occur due to several abnormalities in the image taking machine or due to any human error.

The features such as contrast, entropy, homogeneity, variance etc. popularly known as GLCM features are extracted from the image.

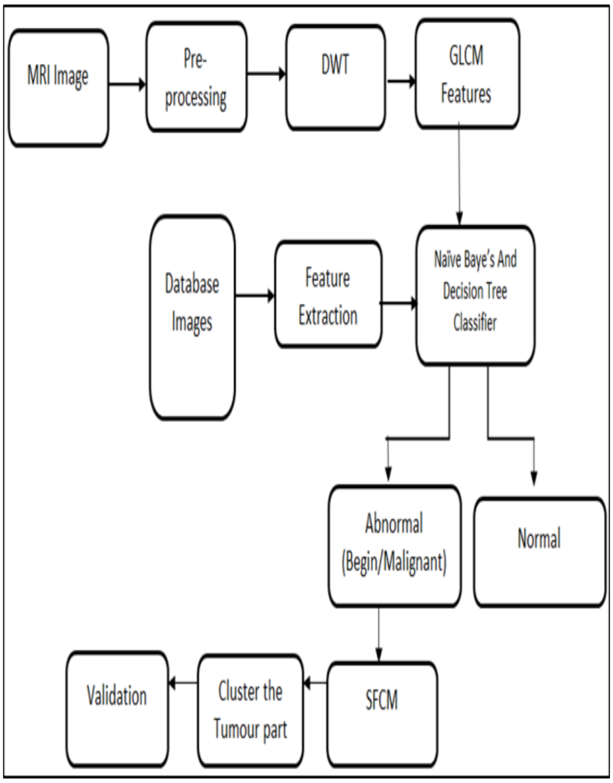

Fig.2 Architecture Diagram

The images of the pancreatic cancer causing tumor are present in the database. A large number of images are taken and the system is already trained on these set of images and the features resembling tumor are set as the threshold units in the machine.

The back propagation network based classifier is used to classify the images. Once the images are fed into the system for classification , a classifier based on the naïve bayes and decision tree algorithm is used for the comparison of MRI images with the previously stored pancreatic MRI images of the patient having pancreatic cancer and tries to find out the features related to tumor.

If the features match then the image is classified as abnormal otherwise the image is normal.

The normal image does not any pancreatic cancer feature whereas the abnormal image is malignant.

The abnormal image now undergoes spatial Fuzzy C Means clustering which now clusters the tumor part of the image and detects the presence of the pancreatic cancer.

V. IMPLEMENTATION

A. Noise Removal

The first step in the detection of the pancreatic cancer is the preprocessing of image. The noise present in the image removed with the application of discrete wavelet transform.

Let us assume that we are having a signal S. The special low pass filter D1 and high pass filter A1 is used for purpose of filtration as shown in the fig. 3 to yield low-pass and high-pass sub bands. After the completion of filtration process about fifty percent of the samples are rejected on accordance with the hyquist criterion..

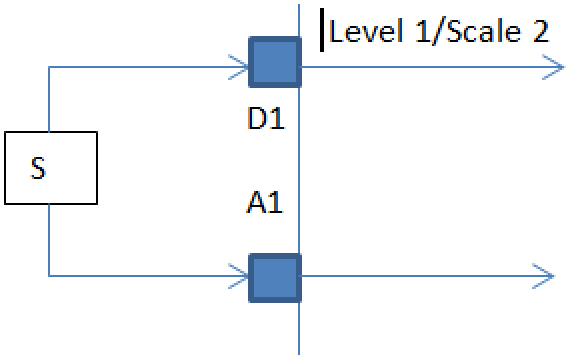

Fig. 3 Filtering of Image

The signals are filtered again and again to yield narrower sub bands. The sub band A1 and sub band D1 is filtered again to yield bands A2 and D2 and so on. The coefficient's length in further sub bands is half to that of the preceding one. In this way it helps to de noise and compress signals

Steps involved

i. Perform a multilevel decomposition which is used to obtain detail coefficients and approximation.

Low Pass sub band and high pass sub bands are termed as approximation level and detail coefficient respectively.

ii. Detailed Analysis of the details obtained based on the analysis helps in finding a suitable threshold technique.

iii. The detail coefficient should be threshold and signal should be reconstructed.

Thresholding operations can be classified to two Soft Thresholding: Coefficient whose value less than threshold value is made to zero and those of greater value are reduced towards zero. It is achieved when we minus the threshold from coefficient value Hard Thresholding: Coefficient with magnitude less than threshold is set to zero whereas coefficient greater than threshold are left unchanged.

B. Features Extraction

i. Angular Second Momentum

$$ASM = \sum_I \sum_j p_{ij}^2$$

$$Energy = \sqrt{ASM}$$

ii. Entropy

$$Entropy = \sum_I \sum_j p_{ij} \log_2 p_{ij}$$

iii. Contrast

$$Contrast = \sum_I \sum_j (i - j)^2 p_{ij}$$

iv. Homogeneity

$$hom = \sum_I \sum_j (1/1+(i-j)^2) p_{ij}$$

v. Variance

$$Var = \sum_i \sum_j (i - \mu)^2 p_{ij}$$

Where the mean of p_{ij} represented as μ

vi. Dissimilarity

$$Dissimilarity = \sum_i \sum_j | i - j | p_{ij}$$

C. Classification

It is based on two algorithms Naïve Bayes theorem and Decision Tree algorithm. Bayes theorem is a method to detect the conditional probability. It gives a natural probability of event information about the tests. When The product of likelihood ratio and prior probability is equal to the value of posterior probability it is called as Bayes theorem.

Steps

i. Frequency table is created using each set of the data.

ii. Likelihood table for each frequency table.

iii. Handling the data: Data is loaded from the CSV file and spread into training and tested assets.

iv. Summarization of data: The properties are summarized in the training data sets to calculate probabilities and make predictions.

The summaries of the datasets is used to generate a single prediction and after that prediction is generated for a given test dataset and a summarized data set. Accuracy is evaluated by the predictions made for a test dataset.

Decision Tree Algorithm: The nodes of the decision tree gets the input in the form of row. The training data set is received by the root node. The decision tree is programmed in such a way that each node will ask true false question related to one of the feature and based on the answer given data is split or partitioned into two subsets the one for which question is true and other one for everything else.

Gini Impurity: It is used to detect the level of uncertainty at a particular node. It ranges between 0 and 1. The lower the value of the gini impurity lower will be the level of uncertainty.

Information Gain: It measures the degree of uncertainty reduced by a question. For each question uncertainty is calculated at starting set. For each question asked data is partitioned and uncertainty of child node is calculated. Weighted average of the uncertainty is taken. More care is taken about a large set with low uncertainty than the small set with high. It is subtracted from the starting uncertainty.

The image is classified into malignant and normal. The malignant image gets clustered and tumor detection is done . The tumor detection validates the presence of pancreatic cancer in the patient.

VI. RESULT

The Fuzzy c means clustering method is used to cluster the tumor part form the malignant image and confirms the presence of pancreatic cancer. The accuracy of the model is detected to be around seventy percent. The model was trained on the seventy five percentages of the data collected from hospitals and rest of the twenty five percent data was used for prediction.

VII. CONCLUSION AND FUTURE WORK

The conclusion of the paper is that Naïve Bayes algorithm and decision tree algorithm is used to predict the pancreatic cancer. Both them were used as classifier which plays the most important role I this model. The classifier classifies the MRI image to malignant and normal one. The accuracy of model is 70 %. Our aim is to achieve 100% accuracy in the model by training the model with variable datasets and with increase in experience.

REFERENCES

1. Jeenal Shah, Sunil Surve, Varsha Turkar "Pancreatic Tumor Detection Using Image Processing" Fr. Conceicao Rodrigues College of Engineering Bandra, Vidyalankar Institute of Technology, 2015
2. K.sujatha, Ponmagal.R. S, Yasoda. K, M. Anand, V. Karthikeyan, V. Srividhya, N.P.G. Bhavani, Su-Qun Cao " Detection Of Pancreatic Tumor Using Bacterial Foraging Algorithm, 2019

3. James J. Farrell, Kimberly Johung, Ying Liang, “Pancreatic Cancer Prediction Through Artificial Neural Network” Department of Therapeutic Radiology,School of Medicine, Yale University, New Haven,CT, United States, 2019
4. Guoda Lian, Chenchen Qian, Shaojie Chen, Jiajia Li and Kaihong Huang “Epidemiology, Detection, and Management of Pancreatic Cancer”, Gastroenterology Research and Practice, 2016
5. Prashant Rawla, Tagore Sunkara, Vinaya Gaduputi”Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors” 2019
6. T.H. Feiroz khan, N.Noor Alleema, Narendra Yadav, Sameer Mishra, Anshuman Shahi "Text Document Clustering using K-Means and Dbscan by using Machine Learning",International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019
7. R.Veeramani,Dr.R.Madhan Mohan, "Iot Based Speech Recognition Controlled Car using Arduino", International Journal of Engineering and Advanced Technology,Volume-9 Issue-1, October 2019
8. S.Babeetha, B. Muruganantham, S. Ganesh Kumar, A. Murugan, "An enhanced kernel weighted collaborative recommended system to alleviate sparsity", International Journal of Electrical and Computer Engineering (IJECE), Volume 10, February 2020, Page No. 447-454
9. R.M.Rani, Dr.M. Pushpalatha, “Generation of Frequent sensor epochs using efficient Parallel Distributed mining algorithm in large IOT”, Computer Communications, Volume 148, 15 December 2019, Pages 107-114
10. S.Babeetha, B. Muruganantham, S. Ganesh Kumar, A. Murugan, "An enhanced kernel weighted collaborative recommended system to alleviate sparsity", International Journal of Electrical and Computer Engineering (IJECE), Volume 10, February 2020, Page No. 447-454

AUTHORS PROFILE

R. Veeramani, Working as an Assistant Professor at Dept. of IT, SRM IST, Chennai, Tamil Nadu, India. His research interests are in Ad-hoc Network, IOT, Cloud Computing and Artificial Intelligent. In addition, he is member of Indian Science Congress Association.

Aryan Goswami, is currently pursuing Bachelor of Technology in Information Technology from SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Harsh Aditya, is currently pursuing Bachelor of Technology in Information Technology from SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Praveen Ranjan, is currently pursuing Bachelor of Technology in Information Technology from SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

.