Detection of Cancerous cells in lung using Machine learning

Capstone Project 2

Submitted in partial fulfilment of the requirement for the award of degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By :

Group No- BT3067

Aryan baliyan(19SCSE1010038)

Rahul Parihar(19SCSE1010059)

SCHOOL OF COMPUTING SCIENCE & ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA

Acknowledgement

I would like to express my special thanks of gratitude to my guide **Mr Praveen Mishra.**

Secondly, I would also like to thank my group members who helped me a lot in finalizing this project within the limited time frame.

NAME: Aryan Baliyan

Rahul Parihar

Declaration

I the undersigned solemnly declare that the project report **Detection of Cancerous cells in lung using Machine learning** is based on my own work carried out during the course of our study under the supervision of our guide MR. Praveen Mishra.

I assert the statements made and conclusions drawn are an outcome of my research work.

I further certify that

The work contained in the report is original and has been done by me under the general supervision of my supervisor.

The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.

We have followed the guidelines provided by the university in writing the report.

Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

INDEX

ABSTRACT

Lung cancer is one of the dangerous and life taking disease in the world. However, early diagnosis and treatment can save life. Although, CT scan imaging is best imaging technique in medical field, it is difficult for doctors to interpret and identify the cancer from CT scan images. Therefore computer aided diagnosis can be helpful for doctors to identify the cancerous cells accurately. Many computer aided techniques using image processing and machine learning has been researched and implemented. The main aim of this research is to evaluate the various computer-aided techniques, analyzing the current best technique and finding out their limitation and drawbacks and finally proposing the new model with improvements in the current best model. The method used was that lung cancer detection techniques were sorted and listed on the basis of their detection accuracy. The techniques were analyzed on each step and overall limitation, drawbacks were pointed out. It is found that some has low accuracy and some has higher accuracy but not nearer to 100%. Therefore, our research targets to increase the accuracy towards 100%.

The development of the computer-aided detection system placed an important role in the clinical analysis for making the decision about the human disease. Among the various disease examination processes, lung cancer needs more attention because it affects both men and women, which leads to increase the mortality rate. Traditional lung cancer prediction techniques failed to manage the accuracy because of low-quality image that affects the segmentation process. So, in this paper new optimized image processing and machine learning technique is introduced to predict the lung cancer. For recognizing lung cancer, non-small cell lung cancer CT scan dataset images are collected. The gathered images are

examined by applying the multilevel brightness-preserving approach which effectively examines each pixel, eliminates the noise and also increase the quality of the lung image. From the noise-removed lung CT image, affected region is segmented by using improved deep neural network that segments region in terms of using layers of network and various features are extracted. Then the effective features are selected with the help of hybrid spiral optimization intelligent-generalized rough set approach, and those features are classified using ensemble classifier. The discussed method increases the lung cancer prediction rate which is examined using MATLAB-based results such as logarithmic loss, mean absolute error, precision, recall and F-score. Finally, we also propose a retraining technique that allows us to overcome difficulties associated to the image labels imbalance. We found that this type of model easily first reduce noise in an image, balances the receptive field size effect, affords more representative features, and easily adaptable to the inconsistency among nodule shape and size. Our intensive experimental results achieved competitive results.

Introduction

Lung cancer is one of the causes of cancer deaths. It is difficult to detect because it arises and shows symptoms in final stage. However, mortality rate and probability can be reduced by early detection and treatment of the disease. Best imaging technique CT imaging are reliable for lung cancer diagnosis because it can disclose every suspected and unsuspected lung cancer nodules . However, variance of intensity in CT scan images and anatomical structure misjudgment by doctors and radiologists might cause difficulty in marking the cancerous cell . Recently, to assist radiologists and doctors detect the cancer accurately computer Aided Diagnosis has become supplement and promising tool . There has been many system developed and research going on detection of lung cancer. However, some systems do not have satisfactory accuracy of detection and some systems still has to be improved to achieve highest accuracy tending to 100%. Image processing techniques and machine learning techniques has been implemented to detect and classify the lung cancer. We studied recent systems developed for cancer detection based on CT scan images of lungs to choose the recent best systems and analysis was conducted on them and new model was proposed.

In the developing technologies, most of the people affected by genetic problem due to the false mutations completely changes human life style. The false mutation entirely changes the structure and function of DNA. The generated wrong mutated DNA cell replaces old DNA cell that creates the abnormal growth of the DNA cell. The abnormal mutation is happened due to the various external factors such as population air breathing, alcohol habits, chemical gas exposure and so on. Mostly, the abnormal cell (DNA) mutation creates tumors that may be occurred in any places such as lung, skin, breast and brain in human body. Among the several tumors, lung cancer is one of the most affected diseases because of the external

factors that generally affect respiratory system. From the study in 2005, the number of deaths is increased up to 159,292 that is increased up to 25% in 2018. From the US report of North American association of central cancer registries, it is declared that 234,030 people are influenced by lung cancer in 2018. Further, the American cancer society conducts the survey in the USA in 2019, 228,150 new peoples are affected by lung cancer in which 111,710 peoples are women and 116,440 are men. From the analysis, 142,670 people died due to the lung cancer (66,020 are women and 76,650 are men). According to the survey, it is finally concluded that lung cancer-affected people ratio is increased gradually in the last 5 years. Based on the analysis, lung cancer is the most common considered diseases in medical field to diagnosis the disease in earlier stage. Normally, the lung cancer is manually predicted with the help of the disease symptoms such as blood coughing, chest pain, shortage of breath, fatigue, weight loss, memory loss, bone fracture, joint pains, headache, neurological problem, bleeding, facial swelling, voice change and change of sputum color. Once the patient has been affected by these technologies, different screening procedures like genetic testing, scopy bronchi, reflex testing, fluid biopsy, biopsy and blood testing have been used continuously for evaluation. From the screening methodologies, national institute for health and care department provides the general guidelines to predict the lung cancer and stages of lung cancer effectively. The discussed screening methodologies successfully examine lung cells and deviations in cells that used to predict the lung cancer, but the prediction accuracy is difficult to sustain. Among the screening methodologies, computerized tomography is one of the effective screening processes that effectively examines the deviations and changes present in human body that is detected by passing X-rays on human body. The 30-min passing of X-rays successfully examines internal organ function, and tissues and affected part details are collected effectively compared to PET and MRI screening

process. By using CT images, automatic lung cancer prediction system is created to detect the disease by using several traditional steps such as image noise removal, region segmentation, cancer feature extraction, feature selection and cancer classification . From the discussed step, region segmentation and highlight selection process is one of the crucial roles because the successful prediction-affected region determines the deviation in normal and cancer cell effectively. Moreover, the segmented region helps to extract the meaningful cancer features which reduce the complexity of the system. Then the feature selection process minimizes computation time to predict the cancer that also reduces the involvement of data overfitting. There are several segmentation techniques such as k-means clustering, distributed clustering, canny edge detection, sobel detection, fuzzy c-means, fuzzy k-means clustering, self-organizing map and Hopfield neural network that are used to extract meaningful region from the captured x-ray images. After that, different features are obtained from the extracted region and with the assistance of different feature selection techniques optimal features are selected such as wrapper methods, ant colony approach, particle swarm optimization, genetic algorithm, fireflies and bacterial swarm optimization that are used to select effective features from set of features. These selected lung cancer features are classified using defined classifiers such as K-nearest neighboring, support vector machine and other intelligent classifiers. Although the traditional automatic system successfully predicts the lung cancer, they still handle the recognition accuracy and also consume more time to process large volume of data. Further the system fails to process minimum quality of CT images that may cause to false lung feature that leads to create more misclassification rate . Then the different authors proposed their opinion about the lung cancer detection process because their thoughts help to get the idea for developing intelligent cancer prediction system. In , detecting lung cancer from computer tomography screening process uses different optimization

algorithms. During the analysis, process 5 approaches median clustering, mean clustering, particle swarm optimization and convergence particle approaches that are used to examine the tumor present in the lung CT image. The captured CT image noise is removed using adaptive median filter, and the histogram analysis is applied to improve the image appearances. Then the different features are extracted, and the affected regions are identified using above-mentioned algorithm. Thus, the author-introduced system effectively recognizes the lung cancer up to 95.89%. In, recognizing CT image lung cancer is with the assistance of the approach to convolution. First, the CT images are collected by stack encoder from the LIDC IDRI dataset. The network extracts various features that are learned using deep neural network. The successful utilization of multiple layers recognizes the abnormal features up to 84.2% of accuracy. Even though several techniques are used, still misclassification and large dimension of data handling are crucial issues. For overcoming above discussed problems in this paper, introduce the intelligent techniques to resolve and improve overall lung cancer detection process. The captured CT lung images are examined continuously using multilevel brightness-preserving approach that eliminates noise from image and enhance quality of the lung image. Due to the importance of segmentation process, in this work enhanced deep neural network approach utilizes multiple layers to extract the affected region effectively. From the derived region, effective and optimized features are selected using hybrid spiral optimization intelligent-generalized rough set approach and those features are classified using ensemble classifier. Finally, the defined intelligent technique-based cancer detection system is developed using MATLAB tool and efficiency of the system is determined using different performance metrics. The rest of the paper is structured as follows on the basis of the above analysis. Section 2 discusses the development procedure of improved deep neural network and ensemble classifier-based lung cancer detection process. Section 3

analyzes the efficiency of improved deep neural network and ensemble classifier and

Literature Reviews/Comparative study-

Several researchers have proposed and implemented detection of lung cancer using different approaches of image processing and machine learning. Aggarwal, Furquan and Kalra proposed a model that provides classification between nodules and normal lung anatomy structure. The method extracts geometrical, statistical and gray level characteristics. LDA is used as classifier and optimal thresholding for segmentation. The system has 84% accuracy, 97.14% sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable. No any machine learning techniques has been used to classify and simple segmentation techniques is used. Therefore, combination of any of its steps in our new model does not provide probability of improvement. Jin, Zhang and Jin used convolution neural network as classifier in his CAD system to detect the lung cancer. The system has 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in Region of interest (ROI) extraction phase which reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy. Sangamithraa and Govindarajan uses K mean unsupervised learning algorithm for clustering or segmentation. It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using grey-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image pre-processing median filter is used for noise removal which can be useful for our new model to remove the noise and improve the accuracy.

The primary goal of automatic lung cancer detection methods is to aid a professional to give better decision during diagnosis. It minimizes energy and time of an expert, and reduces cost of a patient. The investigation made by, indicates that the number of researches conducted on lung cancer detection has rapidly increasing in the recent time. This observation highlights automatic lung cancer detection methods are very significant and it is still a task in progress. Several medical image detection approaches have been proposed by researchers in general and for lung cancer detection in particular. We categorize them into two major approaches. These are discriminative models and handcrafted models. Discriminative models construct detail information about lung anatomy. It digs out low level features and automatically models the relation of original features and its corresponding label. On the other side, different from the discriminative methods, the hand-crafted methods are immensely focus on domain specific understanding about the identification of labels. Nodule presence is challenging to describe, and previous handcrafted methods commonly mis screening the true labels, i.e., the predicted label is not akin with the ground truth. Since a comprehensive analysis is away from the range of our work, we only give some current discriminative methods. Majority of an image detection models are focused on two-dimensional image. For instance, faster-RCNN was introduced by, where the proposed model suggested some bounding boxes during an initial stage and the class decision estimated during second stage. Moreover, the current methods extend to a single stage, where class probabilities as well as bounding boxes are forecasted immediately or without generating the proposal, the class probabilities can be predictable for default boxes. Generally, single stage methods are faster nonetheless two-step methods are better perfect. Recently, proposed a new partially supervised training paradigm, together with a novel weight transfer function, that enables training instance segmentation models on a large set of

categories all of which have box annotations, but only a small fraction of which have mask annotations. Following this, the new batch normalization technique called group normalization is incorporated in deep learning and for example employed for image detection in cascaded way. Very sooner, medical image detection model based on three dimensional convolutional networks were also introduced. Considering small sample size, proposed fast Caps Net for lung cancer screening. The method achieved better result than the tradition CNN and three times faster as well. Taking a computed tomography scan lung image, proposed three-dimensional convolutional network and proposed a multi path CNN for lung cancer detection to detect Cancer and have obtained reasonably encouraging results. However, the time needed to train this model is very expensive and the memory need to process is very huge. Assume how it is so challenging if the task is to be daily work. In addition, these approaches do not take into consideration those previously elaborated challenges and never perform the detection task from the denoised image, but in this paper, we address those challenges.

MODEL

Lung Cancer Detection using CT Scan Images

2017

Research paper model

Image Preprocessing

image pre-processing median filter is used for preprocessing. To remove noise from the image we use median filter and then we use Gaussian filter to implement. This help in smoothening of the image.

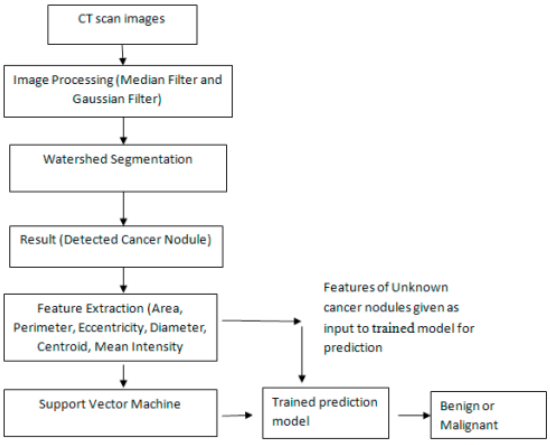
Segmentation-

Watershed method is used for segmentaion. The main feature for that is that is separate and identifies the touching objects in the image. This feature helps in proper segmentation of cancer nodules if it is touching to other false nodules.

Feature extraction

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

Classification-Support Vector Machine (SVM) is used for classification. It is machine learning Algorithm that is supervised. It separates data in two classes .



DATASET = Lung Image Database Consortium (LIDC) archive

Accuracy=92.0%

Sensitivity = 100%

Specificity = 50%

SECOND IMPLEMENTATION

Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier

Image Preprocessing

The introduced method effectively examines each image and its pixel for enhancing the quality of image effectively. During this process, the method enhances the image intensity by computing the mean value of the pixels in the image. Once the image brightness is lower to the neighboring pixels, the pixel value is replaced by utilizing the mean value of pixel. While doing this process, image is divided or partition into different sub-images, each sub-image is analyzed separately and normalizes the image effectively.

Segmentation-

The second step is to extract the affected region from noise-removed lung CT image.

the segmentation is done by applying improved deep neural network (IDNN) because it utilizes the multiple layers while processing the image. Moreover, the network has large volume data that are collected from previous analysis which helps to identify the affected region successfully that also consumes minimum computation time. During the sematic segmentation analysis, the neural network follows several steps such as pixel classification because it effectively examines each pixel and predicts whole inputs present in the lung CT image. After classifying the pixel, localization step is applied to the image for predicting the more information because this information helps to determine normal and abnormal pixel details. And finally, semantic segmentation in which each pixel is labeled correctly whether the pixel belongs to normal and abnormal region. The deep network -based segmentation process includes several steps such as loading the pre-trained network for making segmentation process, selecting input lung cancer image from the defined lung dataset, pixel-labeled (normal and abnormal) images are load and classes are defined and plot the segmentation images using trained network information. Based on the mentioned steps, initially neural network structure is defined with number of layers; in this segmentation process, multiple hidden layers are used to analyze the input lung CT images. After that, weights value is defined for every node for improving overall segmentation accuracy. According to the discussion, in this work VGG deep learning network structure is used to lung image segmentation process.

Feature extraction

During the extraction process, various features such as variance, third-moment skewness, entropy, mean, standard deviation and fourth-moment kurtosis are extracted.

The defined features are extracted from segmented region, in which the each patient has several perceptive of images; these features are derived from entire captured image.

Lung feature selection-

The lung feature selection is done with the help of hybrid intelligent spiral optimization-based generalized rough set approach. The introduced method utilizes minimum control variable, fast selection result, local searching process, simple structure, easiest optimization process and hidden patterns that are easy to identify, easy to make decision effectively, easy to understand and easy to interpret with the results. Due to these advantages, in this work hybrid intelligent spiral optimization-based generalized rough set approach is used to select optimized features from selected features.

DATASET = the cancer imaging archive (CIA) dataset.

Accuracy= 96.2%

Sensitivity = 100%

Specificity = 98.4%

Third implementation

Lung Cancer detection from denoised CT scan image using deep learning

Image Preprocessing

DR-NET is used to denoise lung CT scan images before these images are used to be processed by the detection part of the model. First, the model is trained with LUNA 16 dataset, and then the model is evaluated with KDSB dataset. Both datasets are CT scan images of lung, they have similar characteristics except their Cancer location and Cancer status, where LUNA 16 has Cancer location but no Cancer status and KDSB has Cancer status but no Cancer location. The one we used for detection is the denoised one, i.e., the denoised KDSB images.

Segmentation-

a two-path CNN that consider different receptive field sizes correspondingly. Each path is with unique kernel size, where we named these two-paths as, first path and second path.

Furthermore, in the proposed model we incorporate various feature fusion strategy. For both paths (first and second path), apart from the traditional concatenation approach, we have introduced an efficient way to concatenate their feature map from the fourth and third convolutional layer respectively. First, we transform the fourth convolutional layer of the first path and the third convolutional layer of the second path features and then, after this transformation, we concatenate them together

Feature fusion

In two-path CNN In this sub-section, since the concatenation layer is mainly focuses on discriminant correlation analysis (DCA), we provide the details of DCA. Several feature fusion mechanism has been exploited in convolutional network aiming at obtaining more

DATASET = the cancer imaging archive (CIA) dataset.

Accuracy= 96.66%

Sensitivity = 100%

Specificity = 89.1%

Fourth implementation

A hybrid algorithm for lung cancer classification using SVM and Neural Network 2020 Materials and methods

The proposed classification algorithm KASC is divided into three blocks:

a. Data preprocessing represented by BLOCK-PP (Preprocessing)

b. Feature extraction and optimization is represented by BLOCK-FEO (Feature Extraction and Optimization)

c. Hybridization of SVM and NN for the prediction is presented by BLOCK-HB (Hybridization)

. **Data processing** The selection of the appropriate feature and anticipation with the classifier is the major challenge in most of the CAD models. To handle this situation KASC follows a block model whose flow diagram is given in Fig. The images are stored into a repository that contains two classes namely, “Benign” and “Malignant” and referred as c1 and c2, respectively. The 75% of c1 and c2 dataset has been provided to Block-PP for training purpose and rest 25% of the dataset has been initialized for testing purpose. The BlockPP performs the segmentation of the image using morphologic operations supported by Fuzzy Logic followed by the Region of Interest (ROI) extraction. The extracted ROI for c1 and c2 have been

passed to Block FE-O for feature extraction and optimization of valid feature by applying efficient SURF technique and GA algorithm.

Method

BLOCK HB In this block, SVM technique has been implemented in the initial phase to optimize the feature set which is then provided by block FE-O by applying the efficient SURF and GA algorithm. Here, SVM utilizes two separate kernels for classification, namely linear and polynomial. The SVM preferred only the value of features set that are closed to the kernel and clearly. Here, SVM used only polynomial kernel in contrast to linear. SVM property selection for (a) Linear (b) Polynomial kernel. kernel. In this way, SVM reduces the data size complexity by applying the polynomial kernel property and enhances the valid extracted feature set count for better classification. In this context, only the kernel selected support vectors are passed to FBNN which are known for the preprocessing of raw data. Further, FBNN is going to process the selected kernel features which result in reducing the complexity of the hybrid algorithm to a great extent. The ordinal measures of FBNN have been shown. The training data of SVM could not be directly passed to the FBNN and hence the selected kernel value needs to be picked appropriately.

Algorithm used and accuracy

KASC(kernel Attribute selected Classifier) is used.

Average precision- 98.17%

Accuracy – 98.08%

These accuracy are over 500 data sample

Fifth implementation

Prediction of lung tumor types based on protein attributes by machine learning algorithms.

Feature selection

Twelve attributes weighting models applied on FCdb which gave each feature a weight between 0 to 1. Features that gained weight values higher than 0.50 with at least 50% of weighting algorithms regarded as important protein features. it showed the most important protein attributes selected by more than 50 percent of attribute weighting algorithms (Information gain, Information gain ratio, Rule, Deviation, Chi Squared, Gini index, Uncertainty, Relief, SVM and PCA). Dispersions of features' weight values by two other weighting models.

Classification and prediction

(i)- When X-validation (ten-fold cron validation applied)

The average accuracy Range-

32.27%(SVM Hyper) to

67.36% (For SVM & SVM Linear).

Lowest accuracy- 30%

Highest accuracy- 81.67%

(ii)- When Kappa index followed same pattern

Lowest accuracy- 6.10%

Highest accuracy – 69.09%

(iii)- When Wrapper Validation application, the average age accuracy ranged from

Lowest- 32.21% to highest – 69.53%.

Dataset using and accuracy

SVM Dataset is used here with

Accuracy- 87.73%

When we used Naïve bays then

Accuracy – 77%.

METHOD

ANN In 2017 Amita Desai et al. adopted a method using an artificial neural network (ANN) classifier to predict lung cancer. Images from Manipal Hospital in Goa, V.M.Salgaocar Hospital and SMRC were used. The cropped image is converted into binary which reduces the computational complexity that arises and also decreases the storage issues. it also prepares the image for further morphological operation. After successful pre-processing feature extraction is done by resizing and applying the HAAR wavelet transform, then GLCM is calculated in different directions extracting 7 features from them. The next seven Horlick features are extracted. Then a forward neural network is fed using a backpropagation algorithm. The training accuracy attained is 96% while for testing was 92%. They also achieved 88.7% sensitivity and 97.1% specificity.

METHOD SEVEN

Data preparation and feature selection Proteins that involved in two types of lung tumours obtained from conversion of gene symbols defined by microarray analysis in the GSEA db., using DAVID server. The list of genes associated with two types of lung tumours and those that were common between them showed in Table 1. Data cleaning in original dataset, 59 records classified as SCLC, 30 records belonged to NSCLC class and 25 other records to COMMON tumour classes. For each record 1497 features computed and after removing duplicate, useless and correlated attributes, the number of protein features for each record decreased to 1089 features (less than 28% removed) and this cleaned dataset named as Final Cleaned database (FCdb). Feature selection Twelve attributes weighting models applied on FCdb which gave each feature a weight between 0 to 1. Features that gained weight values higher than 0.50 with at least 50% of weighting algorithms regarded as important protein features. Figure 1 showed the most important protein attributes selected by more than 50 percent of attribute weighting algorithms (Information gain, Information gain ratio, Rule, Deviation, Chi Squared, Gini index, Uncertainty, Relief, SVM and PCA). Dispersions of features' weight values by two other weighting models (SAM and Maximum Relevance) have illustrated in the Figure 2 and Figure 3. Classification and prediction Support vector machine approach Gained accuracies and Kappa values for each SVM model (while Gamma and C set as 0.0065 and 10, respectively and ran with X-validation approach) on 13 datasets (FCdb and 12 datasets that obtained from attribute weighting application: Information gain, Information gain ratio, Rule, Deviation, Chi Squared, Gini index, Uncertainty, Relief, SVM, PCA, SAM and MR) illustrated in the Table 2. Furthermore, Table 3 shows the results of running seven SVM and wrapper validation methods on datasets that derived from attribute weighting (this model cannot be applied on main dataset, FCdb, as required attribute weighted datasets). When X-validation (ten-fold cross validation) applied,

the average accuracy ranged from 32.27% (SVM Hyper) to 67.36% (for SVM and SVM Linear), while the lowest and highest accuracies accounted for the same algorithms (30.0% and 81.67%, respectively). The Kappa index was followed the same pattern, the lowest came from SVM Hyper (−6.10%) and the highest from SVM and SVM Liner (69.09%). With Wrapper validation application, the average accuracies ranged from 33.21% (for SVM Hyper) to 69.53% (SVM) and the minimum and maximum accuracies (23.86% and 71.97%) were again for the same models, respectively.

In this context, FFBPNN considers three regression value including training, validation and test value for better classification. The value of R has been calculated by computing the average of all three regression values. The training regression value has been generated by the passed feature value and the propagated feature value. The FFBPNN creates some validation rules as per the Levenberg architecture and calculates the value of R over the supplied feature value which comes after the validation process. FFBPNN also randomly selects some data values as test data and observe the impact of the test data on supplied input value by calculating its R value. In this work, KASC the hybrid architecture has been tested by varying the number of neurons with different values of regression. Additionally, it has been observed that for the most optimal pre-processing value of R can be achieved by the selected neuron count 15 and it is clearly tabularized in Table 3. Here, the hybrid architecture (KASC) uses neuron count 15 for the selected dataset. The same architecture has been implemented on the test data when the classification process performed on it. Finally, based on this hybrid classification architecture, total classification accuracy, true positive and false positive rate has been calculated.

Model strengths

In 2019, Sanjukta Rani Jena et al. proposed a method that focuses on texture analysis based on feature extraction of images and then classifying them. In image pre-processing, several filters are used to remove the unnecessary noise and stabilize the image. In the feature extraction part, shaped based FETs (Area, Perimeter, Median, Mean, and Variance) and intensity-based FETs (Contrast, Uniformity, Homogeneity) are used. Then the local binary pattern (LBP) is used for texture matching. The performance of LBP is better than other available textual patterns. Then an SVM classifier is used for classification. A hyperplane is chosen such that it maximizes the margin (the distance between a few close points and the hyperplane). In 2019 Nidhi S. Nadakarni et al. proposed an automated system for lung cancer detection at an early stage. CT images from the Cancer Image Archive Database were used in DICOM format. These images were then pre-processed using various image enhancement techniques such as Median Filtering, Smoothening, and Contrast Adjustment to remove noise and improve image quality. Further Morphological opening operations were performed after transforming the grayscale image into a binary image for image segmentation. In the feature extraction method features like area, perimeter, and eccentricity (roundness) are evaluated. Using these features classification of images is done into normal and abnormal using SVM supervised learning classifier. The proposed methodology as said by the authors detects cancer in the early stages accurately.

Main strengths of the proposed model are pointed as below:

- Increase in accuracy of cancer nodule detection than the best current model.

- Classifies the detected lung cancer as malignant or benign.
- Removes salt-pepper noises and speckle noise that creates false detection of cancer Together with strength,

Model has some weakness too. They are pointed as below:

- There is increase in the accuracy but still it has not reached to best level i.e. nearer towards 100%
- It classifies the cancer as just malignant or benign but does not classify into different stages like stage I, II, III, IV.

Conclusion

Incidence of lung cancer has significantly increased over the last three decades and has a worrisome increase in developing countries. LDCT has become the gold standard for lung cancer screening after survival benefits seen in NLST and NELSON studies. An effective lung cancer screening program is still a challenge in developing countries despite a high incidence of lung cancer. LDCT could be a good choice for screening, however, high cost of LDCT, large population size to be screened and low success rates of LDCT make it difficult to implement such a program. Also inadequate infrastructure, lack of human resources, low skilled manpower and lack of financial resources add further difficulties in adopting such a program. An Ideal Screening Method for developing countries should be easily and widely available, easy to perform and must be cost effective. High incidence of tuberculosis in developing countries further compounds the problem by adding to false positive cases during screening. There is a need to develop point of care technology for cost effective lung cancer screening in developing countries as lung cancer is going to be a major burden in coming years.

That's why a model based on machine learning which could be more efficient and more accurate at affordable price is what we need to overcome such a disease.

One of the most fatal diseases to have existed is lung cancer. This disease unfortunately is extremely tough to treat after having spread upto an extent or reaching a serious stage. Computer-Aided Detection (CAD) is one of the constantly growing technologies that help detect cancer by feeding in certain inputs containing patient-related information such as scans like CT-Scan, X-Ray, MRI Scan, unusual symptoms in patients or biomarkers, etc. SVM, CNN, ANN, Watershed Segmentation, Image enhancement, Image processing are a few methods used to improve the accuracy and aid the process. For training, the most popular datasets used are LUNA16, Super Bowl Dataset 2016, and LIDC-IDRI. By the means of this review paper, we aim to list out all the major researches that have been done over the past years and can be improved upon to achieve better results.