Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776646

SVM BASED ANALYSIS OF LUNG CANCER CT IMAGES.

Ravi M.V.
Research Scholar, Electrical Engineering, VTU Belagavi., Assistant Professor, Dept of Electronics & Communication Engineering, S.J.C.Institute of Technology , Chickballapur
ravimvreddy86@gmail.com

Shobha B.N
Professor & HoD, Dept of Electronics & Communication Engineering, S.J.C.Institute of Technology , Chickballapur, bnshobha67@gmail.com

ABSTRACT

The major area of research and development is the treatment of Lung cancer which results in maximum death probability. Initial detection of cancer can aid in the complete healing of cancer. There exist various methods for identifying lung cancer in this research work. To predict cancer, many researchers have submitted their research works. Most of these studies briefly explain regarding the typical identification of lung cancer methods that are examined in this research work. Various approaches have been developed to improve the effectiveness of cancer detection. Diagnosis of cancer employs a wide variety of applications such as neural networks, image processing as well as support vector machines which are addressed in this research.

Keywords: CT image,Digital Image Processing, Lung Cancer, Machine Learning, SVM.

1. INTRODUCTION

Cancer: A condition where few cells of the body proliferate uncontrollably as well as spread to other body parts. Since the body of human being consists billions of cells, cancer can grow anywhere in the body. The cells of human being often divide via a phenomenon known as cell growth as well as cell multiplication to develop new cells as is required by the body. New cells replace the old cells when they stop functioning because of aging or damage. Often when this systematic process fails it results in damaged or defective proliferation of cells which should not happen and tumors arise from these cells. Tumors may or may not be malignant [1].

The cancer of lungs is one sort of tumor. Once the cells of the body start to proliferate out of control, then cancer develops [2].

Normal structure and function of the lungs

Lungs are located in the chest and are sponge-like organs. The right side of lung is comprised of three lobes. Left side of lung consist of two lobes. Since left part of the body consists of heart which requires more space hence the left side of lung is relatively small.

When breathing in, air passes through the mouth or nose and then passes through the trachea which is also known as the windpipe to the lungs. Bronchi are tubes that emerge from the windpipe and then enter the lungs before dissecting into small airways. They break into relatively tiny branches called as bronchioles. The small air sacs referred to as alveoli are located at the tip of the bronchioles [3].

When an individual breathe in, the alveoli accept oxygen into the blood and then exhale carbon dioxide. One of the vital functions of the lungs is to inhale oxygen as well as to exhale carbon dioxide. The cancer of Lung generally originates in cells that make up the bronchi and other regions of the lungs like the alveoli or bronchioles [4].

Types of lung cancer:

Non-small cell lung cancer (NSCLC): NSCLC accounts for approximately 80–85% cases of lung cancer. Adenocarcinoma, giant cell cancer as well as squamous cell cancer, are the major subgroups of NSCLC. As their structure, as well as medications, is the same, these subgroups, which originate from various lung cells are categorized as NSCLC.

Adenocarcinoma: The cells that typically release things like mucus are where adenocarcinomas start growing. The cells that would typically release mucus are where adenocarcinomas begin to form. Although it is the most common kind of lung cancer among those who do not smoke and this kind of malignancy primarily affects the people who smoke. Contrast to other forms of lung cancer, this type is more prevalent in women rather than in men and is more likely to impact young folks.

Adenocarcinoma is detected before it is spread and is typically developed in the outer regions of the lung.Individuals with Adenocarcinoma in situ, previously known as bronchioloalveolar cancer is a type of adenocarcinoma have a better prediction when compared to other lung cancer.

Squamous cell carcinoma: Squamous cells are usually flat cells and cover the interior of lungs' airways and are the origin of squamous cell carcinomas. These cells often develop in the middle part of lungs, close to the bronchus, and are usually related to a past of smoking.

Large cell (undifferentiated) cancer: Large cell cancer can appear anywhere in the lung. These cells typically multiply as well as replicates, making therapy more difficult. Large cell neuroendocrine carcinoma (LCNEC) is a subdivision of big cell cancer which is one of the rapidly developing tumors that resembles a small cell lung cancer [5-6].

2. Digital Image Processing in Lung Cancer

A. Image Acquisition–

Obtaining CT scan images of samples from the Database (ACSC) is the initial step. Images are shown as grayscale images and are saved in MATLAB. When contrasted to MRI as well as scan images, the lung CT images possess very less noise. Therefore, to identify the lungs, CT images can be taken. The primary benefits of computer tomography images are improved resolution, reduced noise, as well as reduced distortion. Ten male photos and their CT scans were saved in a database using JPEG/PNG image standards for the research purpose [7].

B. Image pre-processing-

Every image has to undergo several pre-processing steps like enhancement as well as reducing noise.

The most common image processing algorithms are denoising algorithms. A typical RGB image serves as the input. Since MATLAB does not support RGB format, the RGB image is transformed into a greyscale image. White noise, salt, pepper noise, as well as other noises may be observed in the grayscale image. The most critical challenge in processing of image is white noise. With the help of a filter from the retrieved lung image, white noise can be eliminated. [8]

C. Image Enhancement

Enhancement image is a process of enhancing the digitally recorded images quality by modifying the image with MATLAB^{TM}. Image enhancement is described as a means to increase the feature of an image so that the generated picture is superior to the initial one. For instance, changing an image's brightness or darkness or contrast can be done relatively easily. However, there is no underlying theory that explains how human consciousness affects and what constitutes "excellent" picture augmentation. [9]

The purpose of enhancing an image is to make the image look better visually or to present an improved transform image for upcoming automatic processing of the image. Several pictures like those used in aerial, real-world cinematography, satellite, and medical have a lower resolution as well as noise. To improve the quality of the image, intensity must be improved as well as noise must be eliminated. Depending on the application, the augmentation process varies from one field to another. The Gabor filter enhancement method is used during the phase of image enhancement. [10]

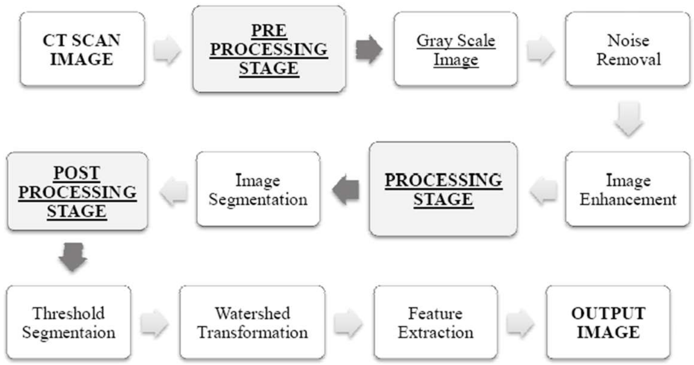

Figure 1: Lung cancer detection and its stages.

D. Processing

The key task at this level is segmentation is described as follows:

i. Image Segmentation

In the vision of computer, segmentation means the division of digital image into several segments and these pixels sets are referred superpixels. Typically, this process is employed to identify objects as well as constraints like lines, curves, and so on, in images. An image can be segmented by giving each pixel a mark so that pixels with the same identity share a set of visual characteristics. Segmenting an image generates a group of segments that represent the entire image or a set of outlines that are obtained directly from the image. Every pixel in an area possesses a certain feature or derived property like color, brightness, or texture. Determining appropriate local features can differentiate an object from other objects and improving the recognition of objects of interest is the overall objective of all image processing procedures. The following step is to examine whether or not each pixel is part of an object's interest. A binary image is generated by this process and is referred to as segmentation. If a pixel is part of an object, it is assigned the value one; if a pixel is not part of an object then it is assigned zero. Upon segmentation, the pixel corresponds to the object. [11]

E. Post-Processing

The following techniques are used for post-processing segmentation.

i. Thresholding approach

To distinguish foreground as well as background, Thresholding is used. The binary image is converted from grey-level image by choosing a suitable threshold value T. The placement as well as the outline of the objects is finally disclosed in the binary image. Initially obtaining a binary image has the advantage of reducing data complexity as well as making the detection as well as categorization processes easier. A particular threshold is used for converting binary image from a gray-level image. Furthermore, all grey-level values below this threshold and above this threshold are categorized as black i.e, 0 and white i.e., 1 respectively. [8]. The global image threshold is computed using Otsu’s technique with the help of the gray thresh function. Depending on statistical criteria, Otsu's approach selects a threshold. To determine the most effective threshold, Otsu proposed a technique for minimizing the weighted sum of an object as well as background pixels' within-class variations. In terms of bimodal histogram images, this strategy produces acceptable results. [12]

ii. Marker-Controlled Watershed Segmentation

Markers are employed in segmentation process called watershed. An interconnected element that is part of an image is a marker. Both internal as well as exterior markers that are associated with a background, as well as objects of interest, are featured as markers. One of the challenging procedures in image processing is distinguishing striking elements in an image. This issue frequently uses the watershed transform. An image can be segmented into distinct boundaries using the watershed segmentation technique-based marker. The advantage of this segmentation is that it generates a distinct answer for a specific image. Marker watershed process also overcomes the problem of segmentation. [13].

Typically, the gradient of the actual picture is used to calculate the watershed transform. It has several benefits: It provides a inclusive partition of the picture in discrete sections with inadequate contrast. It is a straightforward, intuitive method that is quick and parallelizable. Determining the aspects of a Dicom image must be taken into account to accurately diagnose cancer that occurred in lungs. [14]

iii. Feature Extraction

The extracting feature phase of image uses algorithms as well as approaches to identify and extract possibly different areas or features of an image which is considered as a crucial work in processing techniques of image. The lung region can be segmented and its characteristics can be extracted from it, and further, the identification rule can be created to precisely distinguish the cancer nodules present in the lungs. With the help of these principles for diagnosis, cancer nodules that were incorrectly detected as well as segmented can be diagnosed more accurately. The characteristics employed in diagnostic indicators were examined in this research work. [15] Masking, as well as Binarization, are two methods for estimating the likelihood of lung cancer existence.

Binarization Approach

The technique of binarization mainly relies on there are several black pixels. When compared to the white pixels present in normal images of lung, counting begins with the black pixels for regular as well as abnormal images to obtain a standard and can be employed as a threshold. The image is normal whenever the black pixels count of a newest image is larger than the threshold and it is abnormal whenever the black pixels count is less than the threshold. [16]

Masking Approach

The masking strategy is based on the notion that when the percentage of cancer present increases, the masses emerge as white with associated patches inside ROI (lungs). Solid blue color displays signify a normal situation, whereas RGB mass appearances signify the existence of malignancy.[17]

The organization of the research article is divided into seven different sections. The section 1 consists of a brief introduction regarding different technologies and lung cancer discussed in this paper. Section 2 consists of the survey on the lung cancer. Section 3 consists extensive literature survey done on the current date topic of proposed methodology and the algorithm used. Section 4 dedicated for the proposed methodology A detailed explanation of the same has been discussed in the section, followed by the implementation of the algorithm. The Section 5 consists of results of detained in the algorithm. The different parameters that are verified are mentioned and explained with respect to the proposed novel methodology. The Section 6 concludes the paper with the results and performance, the section 7 for the few references which has resulted in the novelty of the present research paper.

3. LITERATURE SURVEY

The authors in [18] conducted experiments to examine if delays were related to illness stage as well as the adverse outcomes on delays in a rapid general medical diagnostic program for patients who have lung cancer. Thorough analyses of the traits of tumors, structure as well as the multiple exemptions that have taken place have been conducted. An average of 565 patient recovery reports was gathered for this research work .

Overall, 51% of individuals suffered lung growths, whereas the remaining half i.e, 8.5% had several injuries, and 111 individuals i.e., 19.6% had radiological abnormalities that were not considered to be serious. First-line wait times were significantly reduced when it comes to hemoptysis. An RODP was designed to simplify the process during the rule-making process. According to the estimates, the delay between the initial and second lines of treatment is responsible for the majority of patient setbacks.

The researchers in [19] looked into a variety of methods for estimating lung growth. This work included the use of ANN, linear dependency analysis (LDA) as well as image processing to list a few self-organizing maps (SOM). Finally, it is suggested to use support vector machines as a characterization tool. While using machine learning, one can examine data as well as recognize patterns using support vector machines. [20] Developed a method to identify lung growth at the outset of their investigation. This method of data preparation initiates the process of image enhancement.

The stage at which the data sets can be evaluated is when they are ready to be tested using data mining techniques and neural systems, both of which are essential for differentiating between therapeutic approaches. Scholars were capable of accomplishing the expected outcome by categorizing information sets as either cancerous or non-dangerous using back-propagation neural networks (BPNN). While making a diagnosis, medical professionals must determine which stage of cancer will be most useful.

The work in [20] determined as well as validated features associated with cancer progression in lungs and its corresponding pathways using biomarker identification that is based on network and quality set enhancement methodologies. The researchers discovered that the results indicated a wide range of novel as well as surprising features with probable physiological capabilities in smoking that are not suggested by prior results in these domains. To cope with the perceptible validation of smoking as well as to categorize the characteristics linked with lung tumor existence and individuals associated with groups who do not smoke, a network-based technique was developed [21].

It has been demonstrated that a smoking-related six-quality mark can determine the probability of surviving and the danger of lung expansion. Smokers may be able to perceive as well as recognize lung development if such a quality mark is adapted. To explore lung growth, the work in [22] used data mining as well as streamlining techniques to generate results from a variety of data sets.

The patterns exposed in databases are utilized to forecast the progress of disease depending on particular medicinal agents recorded in the databases. Through computed CT scans and by using previously identified computer-aided diagnosis (CAD) methods, the authors in [23] validated the detection of neural system enlargement. The features from the CT scans were linked together and then reconfigured to rebuild the lung. The five as well as six central moments, standard deviation (SD), variance, and kurtosis, were employed to identify cancer in the sample. Feed forward neural networks and backward NN are employed for better classification [23].

The authors in [24] worked on the implementation of wide variety of artificial intelligence (AI) methods for diagnosing illnesses as well as providing prescription for a considerable amount of time. To evaluate data related to breast cancer, an ANN can be employed. Multilayer feedforward NN are identical to ANNs and can be employed to calculate the onset of cancer in lungs by utilizing information from microarrays and from machine learning library of UCI. When setting up the system, the back-propagation rule is used. Set of data with variable hidden layers, hubs interconnecting to the similar dataset and cross-approval allow them to be evaluated against each other.

As a consequence of various hidden layers, it is anticipated that the precision of this framework will increase whenever an UCI event from the data set take place. The NCBI data set's hub as well as hidden layer count keeps growing and increasing the analytical precision. Using a analogous neurological method, one can forecast the condition of patient. Through automated decision system, this is possible. In the recent study by [24], CAD analysis, fuzzy-weighted pre-processing, as well as a counterfeit-resistant acknowledgment system were used. Three stages make up the structure [25].

Only four of the high points in the data set may be analysed using classical component analysis, even though the data set has 57 high points. Before the implementation of the primary classifier, a weighting strategy corresponding to prior handling of fuzzy weighting was used as a basic stage. Finally, to spot the forgeries a classifier for counterfeit safe acknowledgment was used. The lung data set was utilized to assess the researchers' programmed technique for identifying tumors. It was quite promising for future grouping scenarios to discover the characterization as 100% accurate [26].

An prediction model [27] for predicting emotions incorporates user data from social platforms, explicit data rating, as well as sentiment data collected from user reviews can be used to generate ideas. It is feasible to enhance the predictions, ratings, as well as recommendations with the aid of this technology. For minimized nonstop capacity, the authors in [26] built trustworthy prediction algorithms. Even though this technology doesn't offer much efficiency as standard higher-order feedforward systems, this also offers dependable as well as fruitful architecture by retaining its rapid characteristics of learning. The pi-sigma group assumed that they would use a unique type of edge polynomial in their polynomial system of edge. Some multi-variate polynomial has the potential to address RPN and is recognized as a solution.

The research suggests that the system has an effective computation that leads to smooth speculation and ongoing learning. The authors in [28] have developed a strategy for anticipating lung disease by identifying it quickly, and addressing it when the lung cancer is minor. Some peaks were eliminated from the image to enhance readiness for lung cancer. It has been found that the ability to predict pulmonary development is impacted by the acknowledgment-based systems' architecture.

Image processing approaches were used while conducting a thorough evaluation of past knowledge of lung tumor requirements. Computer-based Tools for image pre-processing can assist to predict and regulate lung expansion. Lung cancer that are major can be accurately forecasted as per [29] by utilizing a novel method that can identify biological signals in the body. Since tumours are complicated, it is impossible to predict their current status without incorporating several elements12,600 NSCLC articulation features with excellent levels are used in the study to demonstrate the existence of 9 feature marks that can be used to identify NSCLC lung cancer as well as establish genetic markers for cancer.

The scholars employed a technique based on previously opaque organic marker arrangement to obtain an optimum prediction accuracy of about 99.75% for identifying the disorder and to find genetic markers for NSCLC. Crossover inheritance, as well as molecular swarm optimization strategies, was proposed in [30] and the algorithm of MLPNN was employed to examine the component of CT images. When contrasted with lung CT scans, the above strategy was shown to be a trustworthy source of information.

Pre-processed images were used to help in the extraction procedure for assisting images to reduce agitation for better extraction outcomes. With the help of the MAD technique, the authors were able to extract these features. With the aid of GAPSO, specific features were chosen. The GAPSO-MLPNN algorithm is used to generate the overall image result. [30] A back-propagation neural network (BPNN) was used to build a replica for predicting input costs.

The researcher thereafter suggests a self-evolving trading technique that corresponds with the market regulations for upcoming trading and the results of the testing. At last, the alternative techniques are contrasted with the traditional ones to show how their methodology has changed over time. Extracting of Hybrid feature increases the authenticity of the ECG data is been suggested in [31]. A parallel pattern recognition paradigm for ECGs was devised, which boosted detection capability across various ECG feature spaces. The researchers in [32] have demonstrated a hybrid approach and tested it on genuine Sina Weibodata sets. Figure 1 provides information on the machine learning models employed in this study.

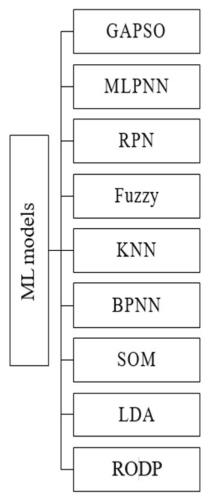

Figure 1: Conventional Model for Machine Learning

4. PROPOSED METHODOLOGY.

An SVM classifier is used in this area to categorize lung cancer patients according to their symptoms. In the suggested technique, pre-processing occurs before data collection. The chosen classifiers are then trained as well as assessed on the reference data set for a second time using a conventional 10-fold cross-validation approach. The information is analysed as well as examined to find out the most useful way of identifying lung cancer. Figure 2 gives a comprehensive summary of the developed strategy.

Data are provided initially and are inputted in sequence. Different types of information from which these inputs are derived are required for various analyses and the data is retrieved according to the outcomes. These pre-processing techniques divide the input data into miniature groups and further effectively handle them. Different sets of data are categorized for this processing step. The fundamental form of the data set as well as the rate of operation is determined and categorized in these segmentation processes. The applications of its toddlers are retrieved after they have been finally classified.

A. Acquisition of Data.

A dataset obtained from the California University, Irvine's online repository called Lung Cancer was utilized for this research work. The dataset as a whole consists of 32 examples, 57 features, and one class attribute. Evaluation as well as comparison of SVM performance among other things is the fundamental objective of the suggested four research works.

B. Preprocessing of Data.

Initial phase of identifying lung cancer is pre-processing, which addresses any differences in the dataset and thereby removes data that are not relevant. To boost the overall accuracy of the data set, missing values are provided using the nearest neighbour technique with three neighbours. Samples for testing, as well as training, are necessary.

C. Training Samples and Testing Samples.

A neural network is a form of AI and is used to develop and evaluate the input data samples first. Beginning with the input data, a random number is used to calculate the weight of the neural network. The neural networks are evaluated using the same dataset that was used for training after being trained a sample dataset. Data are weighted to determine the errors that occur during the process of classification, and further errors are corrected by reweighting the dataset.

D. Feature Extraction.

The data set must be further subdivided to extract a substantial number of attributes that reduces the difficulty of diagnosing lung cancer. The cancer in lungs is formed as a consequence of proliferation of cancer cells will be removed using this detection method. The PSO programming language is utilized in this feature extraction method. In pattern recognition algorithms, feature extraction from input data is utilized to extract the key features that are more valuable as well as non-repeatable, and it is also used to gather cancer-related data to identify patient situations for interpretation as well as to classify data using SVMs.

E. SVM Classification.

The course of arranging data as logical groups is referred to as classification. It is possible to make decisions using both organized as well as unstructured data, which the system is capable of analyzing. By using methods like data mining, malware as well as denial-of-service (DoS) attacks can be avoided as well as controlled. As a result, texts are automatically corrected as well as classified to maintain their authenticity throughout the process. Premalignant cancer has been included as either malignant (M) or benign (B) in the traditional cancer classifications. Individuals in this category will receive a higher degree of care and treatment options. Data security areas including copied, transmitted, as well as extracted data are always on the watch for abnormal behavior. Classification is the method of assigning data depending on the application to improve its functionality as well as accessibility. As a result, redundant information can be reduced as well as both backup and storage costs can be minimized. Processing times may thus be greatly reduced in specific circumstances.

The SVM method trains machines to learn on their own to differentiate between things belonging to different groups. An ideal SVM model would include hyperplanes that segregate classes because a hyperplane's margin is decided by the sample that is nearest to it. The primary function of SVM is to achieve maximum profit margins. The classification is given by using a feature vector of dimension as well as the training data set xi.

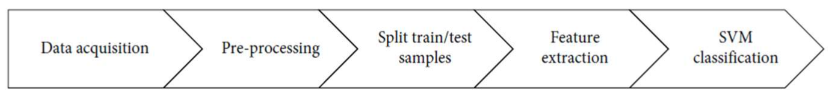

Figure 2: Proposed design

$$f(x) = \text{sign}\ (wx + b) \qquad (1)$$

$$y_i(w^T x_i + b) > 1.,\ i=1,2,\ldots\ldots..N \qquad (2)$$

$$\text{mimimize } 0.5\ w^T\ ws.t.y_i\ (w^T x_i + b) < 1.,\ i=1,2\ldots\ldots N \qquad (3)$$

$$L\ (w,b,r) = 0.5\ w.w^T + \sum_i r_i \left(1 - y_i(w^T x_i + b)\right) \qquad (4)$$

Equation (1) is modified as Equation (2) by adding a slack variable ξ as shown in Equation (3) since non-separable information cannot be segregated. Equation (6) is derived by multiplying 0.5 $\omega T\omega$ by the penalty constant C. A penalty constant C may be assigned if you incorrectly classify a training model. The algorithm chooses a hyperplane with a significant edge for smaller values of C, which could lead to more points and being incorrectly classified. Based on the technique employed, a wide variety of possible values for parameter C are often used, as well as the best possible solution is assessed by employing either a distinct validation set or by using cross-validation to validate the functioning using just one trained dataset.

Feature learning vs. feature selection

Instead of choosing from a list that is already selected or defined CNN approaches construct features from the beginning, in contrast to Radiomic / texture analysis techniques.

Hierarchical features

The network is capable of understanding the connections between the features as the initial layers of a CNN consist of multiple features in a sophisticated manner when compared to that of a feature extraction executed in a single step. Let us assume the following illustration: a textural characteristic such as the local entropy of the combined histogram can be utilized to identify speculations reaching the parenchyma. However, a CNN find out this as well as understand the speculations that cover the entire nodule's circumference and this is a symptom of a cancerous nodule.

End-to-end learning

CNNs are commonly given training "end-to-end," which means the network as a whole is designed to overcome the issue that all network parameters are modified till the peak classification rate is attained. On the other hand, there was no certainty that the entire pipeline would be modified; instead, each stage of the Lung that is based on the texture technique had to be designed and optimized independently.

Segmentation-free

The nodule segmentation step is not necessary for the CNN technique as segmentation is accomplished automatically by the algorithm. Further investigation of our Lung X method as shown in figure 3 revealed the prediction score as a highly sensitive.

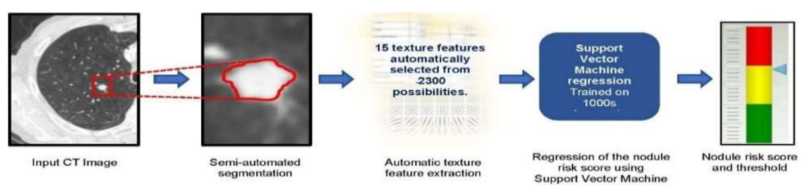

Figure 3: Lung X scanning system block diagram

5. Results and Discussion

A. Sample Results

Figure 4 and figure 5 shows 2 sample results that are obtained from the proposed model. The proposed model process on the CT scan image and extract the lung area. The two sample results shown here consists of Adenocarcinoma Cancer and the normal lung image. From the below image, it is clearly seen that the proposed algorithm can extract The cancer issues part of the lung using CT scan image.

Adenocarcinoma Lung Cancer

Figure 4: The different stages of cancer extraction using the proposed model. (The lung sample Adenocarcinoma Lung Cancer)

Normal Image

Figure 6: the image of a normal lung and the extraction process.

B. Parameters Verified

The seven Haralick features extracted are as follows, where Ng is the number of gray levels, P is the normalized symmetric GLCM and p(I,j) is the (I, j)th element of the normalized GLCM.

$$\text{Energy} = \sum i \sum j \text{ p } (i,j)^2 \quad (5)$$

Energy is calculates the local uniformity of gray levels in an image (as in equation (5)). Higher the similarities in pixels, higher is the energy value (equation(6)).

$$\text{Correlation} = \sum i \sum j \ \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \, ,\sigma_y} \quad (6)$$

Where μ_x, μ_y, $\sigma_{x\ ,\ \sigma_y}$ are the means and standard deviations of the GLCM.

Correlation is the measure of linear dependency of gray intensity values in the co-occurrence matrix.

$$\text{Variance} = \sum i \sum j \ (i-u)^2 \text{p(i,j)} \quad (7)$$

Variance feature measures the spread of intensity values of GLCM pixels about the mean. It is similar to entropy.

$$\text{IDM} = \sum i \sum j \ \frac{1}{1+(i-j)^2} \ \text{ p(i , j)} \quad (8)$$

Inverse difference moment (IDM) gives an account of the local homogeneity in the image. When the local gray level in an image is uniform, IDM is high.

$$\text{Difference Entropy} = \sum_{i=0}^{N_g-1} P_{(x-y)}(i) \log P_{(x-y)}\ (i)) \quad (10)$$

$$\text{IMC1} = \sum \frac{HXY-HXY1}{\max(HX,HY)} \quad (11)$$

IMC1 is the information coefficient of correlation I, where

$$\text{HXY} = -\sum i \sum j \text{ p(i , j) log(p(i,j))} \quad (12)$$

$$\text{HXY1} = -\sum i \sum j \; \text{p(i , j)log} \; \{p_x(i)\; p_y(j)\} \qquad (13)$$

$$\text{HXY1} = -\sum i \sum j \; p_x(i)\; p_y(j) \; \log \; \{p_x(i)\; p_y(j)\} \qquad (11)$$

$$\text{Contrast} = \sum i \sum j \; (i-j)^2 \; \text{P(i, j)} \qquad (12)$$

Contrast indicates the intensity variations between the pixel under consideration and its neighbouring pixel. Larger contrast means larger variation. The total number of features from each image was 252 (7 features * 4 directions * 3 HAAR approximations (horizontal, vertical and diagonal) * 3 different resolutions of image = 252 features. This is shown form equation (6) to (10).

C. Confusion Matrix

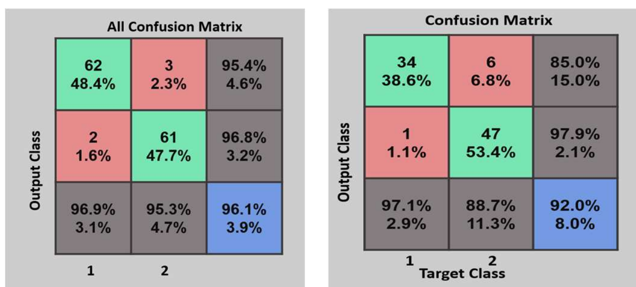

Figure 7: Confusion matrix of the proposed system

The training accuracy was 96% and testing accuracy was 92%. The sensitivity was 88.7% and specificity was calculated to be 97.1%. The confusion matrix of the proposed algorithm is shown in figure 7.

Table 1: Comparison table for accuracy for different models.

Model	Accuracy
Bayes Net	87.06%
Naïve Bayes	89.64%
Decision Tree	90.29%
Random Forest	91.26%
Artificial Neural Network	92.23%

Table 2: Comparison table of different state of art parameters of different model.

Model	TP Rate	FP Rate	Precision	Recall	F- Measure	MCC	ROC Area
Bayes Net	0.871	0.501	0.865	0.871	0.868	0.387	0.855
Naïve Bayes	0.896	0.432	0.89	0.896	0.893	0.5	0.902
Decision Tree	0.903	0.409	0.897	0.903	0.899	0.531	0.787
Random Forest	0.913	0.386	0.907	0.913	0.909	0.574	0.917
Artificial Neural Network	0.922	0.318	0.919	0.922	0.921	0.633	0.938

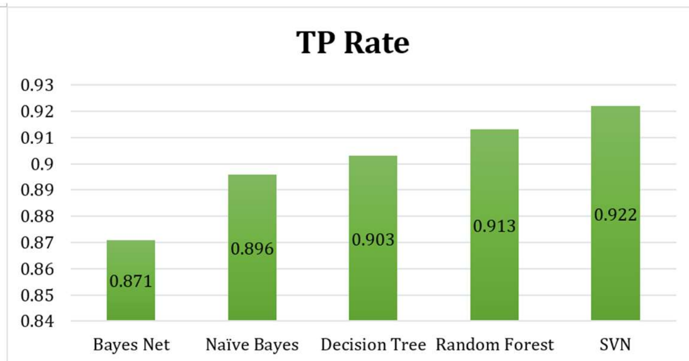

Figure 8: The comparison of different models on SVN with respect to TP rate.

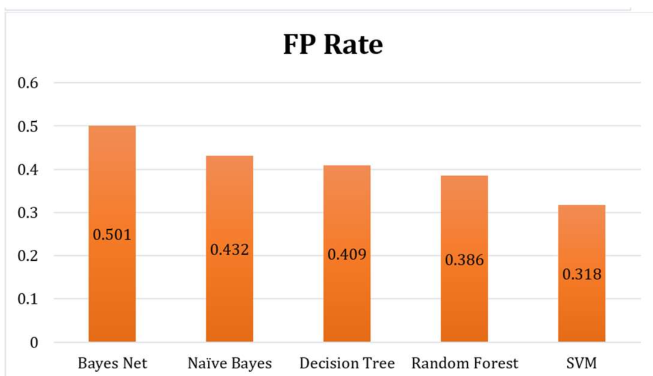

Figure 9: The comparison of FP rate With the different models and SVM.

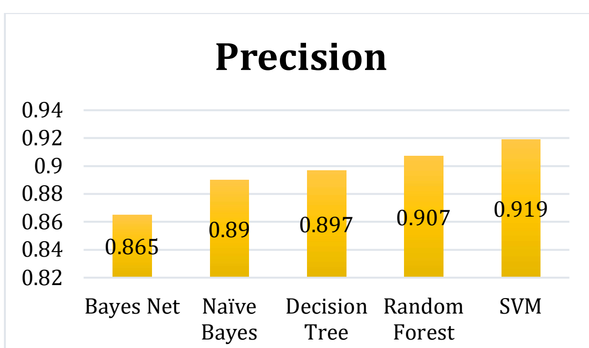

Figure 10: The comparison of precision with respect to different models and SVM.

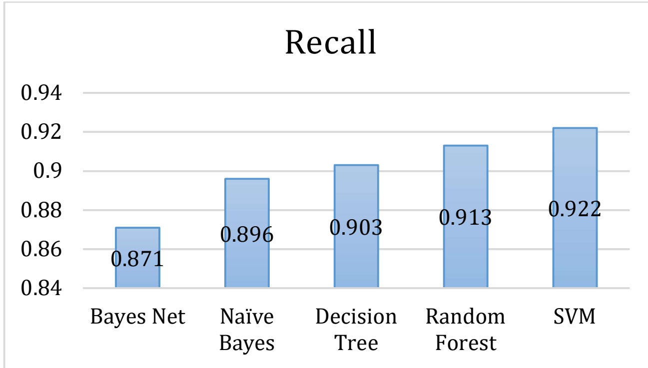

Figure 11: The comparison of SVM with the other models with respect to recall.

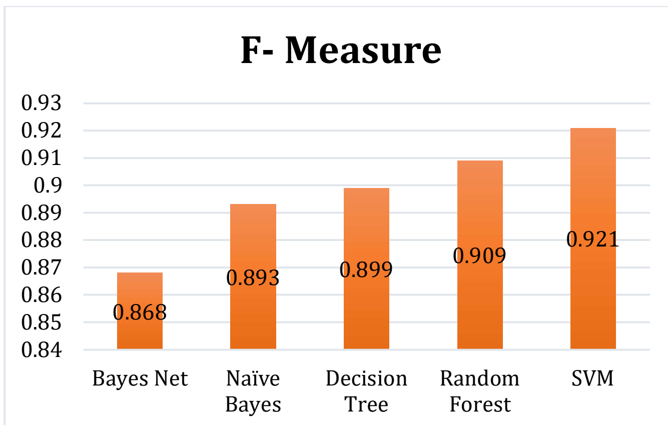

Figure 12: The comparison of SVM with other models With respect to F measure.

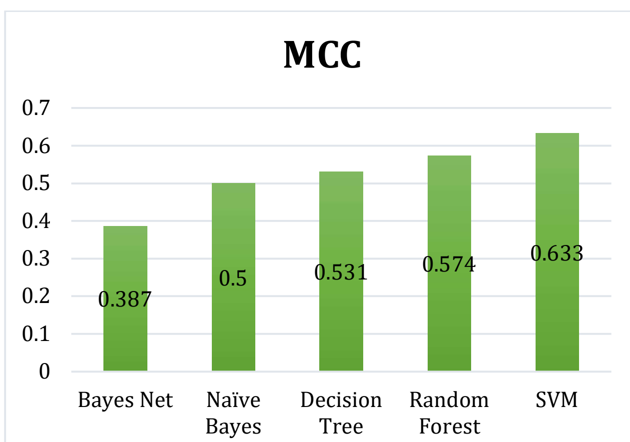

Figure 13: The comparison of SVM with other models with respect to MCC

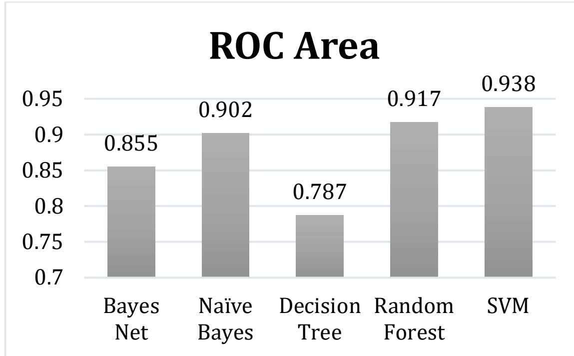

Figure 14: The comparison of SVM with other models with respect to ROC area

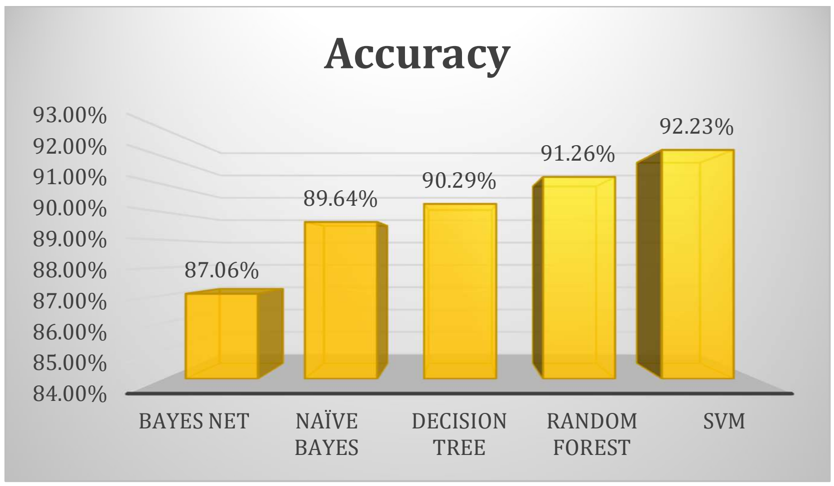

Figure 15: The Comparison of SVM with respect to order models in terms of accuracy.

The table 1, Shows the accuracy obtained for different models. Among which the SVM has the highest accuracy of 92.33%. The table two is the comparison of different state of art parameters, That are very fired during the development of new model with respect to other models. They figure 8, showing the true positive rate of different algorithms. Figure 9 shows the false positive rate come up. Figure 10 shows the precision. Come up. Figure 11 shows the recall, figure 12 shows the F measure, figure 13, showing the MCC, And the last figure 14 showing the ROC area of proposed algorithm compared with existing algorithms. Figure 15 is the comparison bit of accuracy of the proposed algorithm with respect to the existing algorithms.

6. Conclusion.

The majority of tumour cells are covered by one another therefore; initial identification of lung cancer becomes difficult. This research work has examined a wide range of techniques for identifying lung tumours in their early stages. Manual evaluation of samples is time-consuming, imprecise, as well as labour-intensive to avoid diagnostic problems. From the experimental results it is found that when compared to basic textural patterns, the Local binary pattern outperforms as its histogram features were higher than those of the earlier.

7. REFERENCES.

[1] C. Anil Kumar, et. al., " Lung cancer prediction from the text data set using machine learning", Volume 2022, Article ID 6254177, https://doi.org/10.1155/2022/6254177

[2] J. D'Cruz, A. Jadhav, A. Dighe, V. Chavan, and J. Chaudhari, "Detection of lung cancer using backpropagation neural networks and genetic algorithm," Computing Technologies and Applications, vol. 6, pp. 823–827, 2016.

[3] Anil Kumar C, Chethan Venkatesh, Lavanya Vaishnavi D A, Harish S, "Computer vision based Hand gesture recognition system", Neuro Quantology, Jul 2022, Volume 20, Issue 7, Page 2859-2866, DOI: 10.14704/nq.2022.20.7.NQ33365

[4] P. Vivekanandan, "An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data," in In 2013 International conference on information communication and embedded systems (Icices), pp. 761–768, IEEE, USA, February 2013.

[5] Lavanya Vaishnavi D A, Anil Kumar C, Harish S, Divya M L, " MediaPipe to Recognise the Hand Gestures", WSEAS TRANSACTIONS on SIGNAL PROCESSING Volume 18, 2022, DOI: 10.37394/232014.2022.18.19

[6] J. Shen, J. Wu, M. Xu, D. Gan, B. An, and F. Liu, "A hybrid method to predict postoperative survival of lung cancer using improved SMOTE and adaptive SVM," Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 2213194, 15 pages, 2021.

[7] Anil Kumar C, Lavanya Vaishnavi D A, "Renesas Based Smart ROBOT using Embedded System". International Journal of Research in Science, Commerce, Arts, Management and Technology, Volume 2, Issue 1, February 2021. (DOI: 10.48175/IJRSCAMT-101),

[8] S. Mandal and I. Banerjee, "Cancer classification using neural network," International Journal, vol. 172, pp. 18–49, 2015.

[9] Anil Kumar C., Pradeep Kumar B.P, Venu K N, Lavanya Vaishnavi D.A., "Intra Prediction Algorithm for Video Frames of H.264", 2021, Natural Volatiles and essential oils journal, Page 51-58, Volume: 8 Issue: 5.

[10] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," Qubahan Academic Journal, vol. 1, no. 2, pp. 141–149, 2021.

[11] F. Taher, N. Prakash, A. Shaffie, A. Soliman, and A. El-Baz, "An overview of lung cancer classification algorithms and their performances," IAENG International Journal of Computer Science, vol. 48, no. 4, 2021.

[12] Anil Kumar C, Lavanya Vaishnavi D A, "Renesas Based Smart ROBOT using Embedded System", International Journal of Research in Science, Commerce, Arts, Management and Technology, Volume 2, Issue 1, February 2021, ISSN (Online) 2582-8045

[13] Jaweed and F. Siddiqui, "Implementation of machine learning in lung cancer prediction and prognosis: a review," in Cyber Intelligence and Information Retrieval, pp. 225–231, India, 2022

[14] V. N. Jenipher and S. Radhika, "SVM kernel methods with data normalization for lung cancer survivability prediction application," in In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1294–1299, IEEE, Canada, February 2021.

[15] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew, "Prediction of pulmonary diseases with electronic nose using SVM and XGBoost," IEEE Sensors Journal, vol. 21, no. 18, pp. 20886–20895, 2021.

[16] B. R. Manju, V. Athira, and A. Rajendran, "Efficient multi-level lung cancer prediction model using support vector machine classifier," in In IOP Conference Series: Materials Science and Engineering, vol. 1012, India, 2021no. 1, Article ID 012034IOP Publishing.

[17] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A hybrid algorithm for lung cancer classification using SVM and neural networks," ICT Express, vol. 7, no. 3, pp. 335–341, 2021.

[18] G. Yin, Y. Song, X. Li et al., "Prediction of mediastinal lymph node metastasis based on 18F-FDG PET/CT imaging using support vector machine in non-small cell lung cancer," European Radiology, vol. 31, no. 6, pp. 3983–3992, 2021.

[19] Y. Xie, W.-Y. Meng, R.-Z. Li et al., "Early lung cancer diagnostic biomarker discovery by machine learning methods," Translational Oncology, vol. 14, no. 1, article 100907, 2021.

[20] M. Yakar, D. Etiz, M. Metintas, G. Ak, and O. Celik, "Prediction of radiation pneumonitis with machine learning in stage III lung cancer: a pilot study," Technology in Cancer Research & Treatment, vol. 20, p. 153303382110163, 2021.

[21] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," Expert Systems with Applications, vol. 164, article 113981, 2021.

[22] S. Vijh, R. Sarma, and S. Kumar, "Lung tumor segmentation using marker-controlled watershed and support vector machine," International Journal of E-Health and Medical Communications (IJEHMC), vol. 12, no. 2, pp. 51–64, 2021.

[23] S. S. Ashwini, M. Z. Kurian, and M. Nagaraja, "Lung cancer detection and prediction using customized selective segmentation technique with SVM classifier," in Emerging Research in Computing, Information, Communication and Applications, pp. 37–44, Springer, Singapore, 2022.

[24] H. A. Miller, X. Yin, S. A. Smith et al., "Evaluation of disease staging and chemotherapeutic response in non-small cell lung cancer from patient tumor-derived metabolomic data," Lung Cancer, vol. 156, pp. 20–30, 2021.

[25] X. Fang, M. Netzer, C. Baumgartner, C. Bai, and X. Wang, "Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer," Cancer Treatment Reviews, vol. 39, no. 1, pp. 77–88, 2013.

[26] N. L. Guo and Y. W. Wan, "Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival," Artificial Intelligence in Medicine, vol. 55, no. 2, pp. 97–105, 2012.

[27] V. Krishnaiah, G. Narsimha, and N. S. Chandra, “Diagnosis of lung cancer prediction system using data mining classification techniques,” International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39–45, 2013.

[28] J. Kuruvilla and K. Gunavathi, “Lung cancer classification using neural networks for CT images,” Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 202–209, 2014.

[29] K. Polat and S. Güneş, “Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer,” Expert Systems with Applications, vol. 34, no. 1, pp. 214–221, 2008.

[30] Y. Qian, Y. Zhang, X. Ma, H. Yu, and L. Peng, “EARS: emotion-aware recommender system based on hybrid information fusion,” Information Fusion, vol. 46, pp. 141–146, 2019.

[31] X. Fang, M. Netzer, C. Baumgartner, C. Bai, and X. Wang, “Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer,” Cancer Treatment Reviews, vol. 39, no. 1, pp. 77–88, 2013.