DOI 10.4149/BLL_2024_34

LITERATURE SURVEY

Integrated global and local feature extraction and classification from computerized tomography (CT) images for lung cancer classification

Murugaiyan SURESH KUMAR[1], Panneerselvam DEEPAK[2], Parthasarathy VASANTHAN[1], Kandasamy VIJAYAKUMAR[3]

Department of Information Technology, Sri SaiRam Engineering College, Chennai, India.
sureshkumar.it@sairam.edu.in

ABSTRACT
Despite being the second most often diagnosed form of cancer, lung cancers are rarely found in the general population. It is proposed in this study to employ a methodology of extracting both global and local features from CT scan images for the identification of lung cancer.
Data gathering, globalised and localised training as well as testing the model are all part of this structure. This study makes use of 800 CT scan images. Images are pre-processed by warping and cropping in advance of the global testing step. Each image is represented by a feature vector employing eight distinct types of image characteristics, which are taken from the images. After creating feature vectors, three machine learning methods are employed to create detection models. Every medical image has been partitioned over a series of simple divisions throughout the training and testing process locally. To describe each block, feature vectors are derived from the image features that worked effectively in the general phase of the experiment. Similar extracted features are then used to build detection systems for all picture blocks using the learning strategies that were effective in the global stage. SVM using Haar Wavelet characteristics had an accuracy, sensitivity, and specificity of 89%, 90%, and 89%, respectively. One might get 90%-accurate results with SVM and 91%-sensitive and 91%-specific results using SVM plus HOG features. Finally, the utilisation of SVM with Gabor Filter characteristics achieved the greatest correctness, specificity, and sensitivity values, particularly 87%, 86%, and 87%, respectively *(Tab. 3, Fig. 7, Ref. 18)*. Text in PDF *www.elis.sk*

KEY WORDS: feature extraction, support vector machine, lung cancer, classification, machine learning.

Introduction

Cancer is a disease that affects a large percentage of the population and is extremely deadly (1). Within the human body, aberrant cells are growing and spreading. If identified and properly diagnosed, this can be effectively treated. In the regular course of things, damaged cells are updated with new ones. When this mechanism collapses and injured cells are not replaced, cancer develops. These cancerous cells might spread to other parts of the body and form metastases. Cancer that originates in the lungs is known as lung cancer. In terms of frequency, it is ranked second. Only 17.4% of patients in the United States survive without therapy for five years following cancer diagnosis, and that number is far lower in undeveloped nations (2).

It is vital to note that early identification of lung cancer can lead to a faster recovery time as well as reduce the complexity and cost of the therapy. It is possible for a 5-year survival rate in the United States to rise from 20% to 65–70% with early detection and treatment of the disease (3). Blood tests, radiological tests, endoscopic procedures, and biopsies are all options for detecting and diagnosing lung cancer. Some tests have advantages and disadvantages, while others might be used for specific purposes. Scanners use CT (Computerized Tomography) scans to provide an accurate diagnosis quickly and painlessly. They may also get information on the tumour's structure, volume, and position (4).

An x-ray scanner captures several pictures from different angles of the same cranial area, resulting in a CT scan, which creates three-dimensional pictures inside the system. Additionally, a CT scan aids in the diagnosis of medical disorders inside the thorax. Specialists often use imaging tests with a contrast-enhancing agent infused into the bloodstream to identify lung cancer (5). This im-

[1]Department of Information Technology, Sri SaiRam Engineering College, Chennai, India, [2]Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, and [3]Department of Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai, India

Address for correspondence: Murugaiyan SURESH KUMAR, Department of Information Technology, Sri SaiRam Engineering College, Chennai, India.

Abbreviations: CT – Computerized Tomography, CAD – Computer Aided Detection, CNN – Convolution Neural Network, KNN – K Nearest Neighbours, GLSM – Gray-level co-occurrence matrix, DCNN – Deep Convolution Neural Network, SVM – Support Vector Machine, ROI – Region of Interest, ROC – Region of Curve

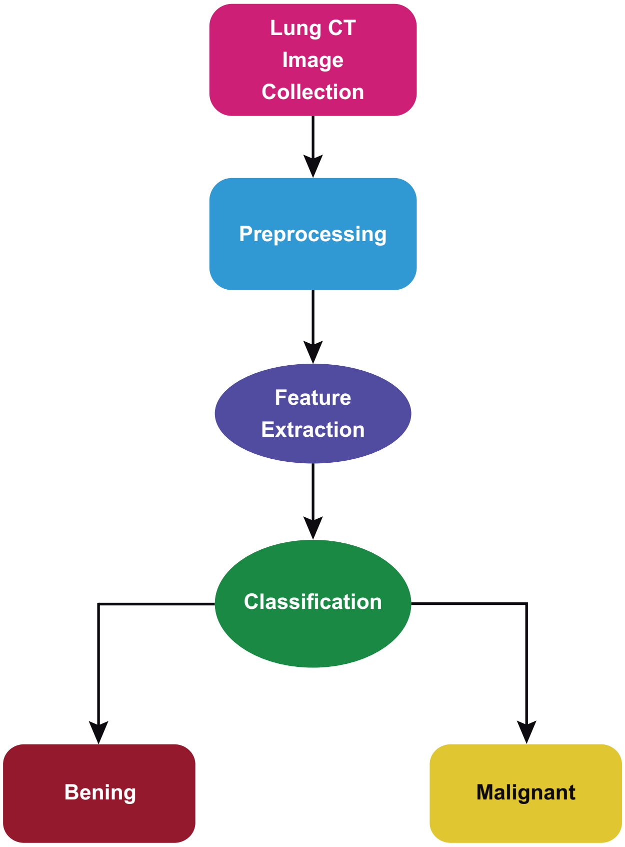

Fig. 1. General structure of lung cancer classification model.

age provides a clearer view of the lung's finer characteristics. To better diagnose lung cancer, CT scans like this one offer highly detailed images of the patient's chest. Radiologists use CAD systems to detect signs of malignancy in their patients' images. For radiograph-based images, such schemes employ numerous image processing techniques as well as machine learning approaches to identify questionable areas (6).

Doctors and specialists benefit from the detection of these zones, which aids in their final evaluation of the images. Many various computer-aided detection (CAD) methods have been proposed by researchers to identify and categorise lung cancer in CT scan images. The following section provides a brief overview of these systems. These systems employ a variety of image attributes and machine learning methods in an effort to identify and classify objects (7). Segmentation, Feature Extraction and Characterization are the three steps in the detection of lung cancer. The establishment of lung and colon cancer diagnoses with the use of Artificial Intelligence is becoming a hot study topic. Lung as well as colon cancer are now the leading causes of cancer-related mortality. General structure of Lung cancer classification Model is depicted in Figure 1.

Lung cancer has the highest incidence and mortality rate of any cancer in the world. Dataset selection over testing and training, image pre-processing with the specified image dataset, extracting the features depending on numerous factors, and image classification utilizing classification methods are the four basic stages in medical image classification. Texture, colour structure, and shape extraction are some of the more traditional methods for extracting feature information. Due to their sensitivity to light intensity, quantization mistakes and computational complexity on representation, these extracted features are fraught with serious problems (8). Other issues include extracting geographical information, indexing and retrieving images. Machine learning may be used to improve the accuracy and precision of classifications because of these issues.

A range of feature extraction, feature optimization, and classification approaches may be used to biomedical images via machine learning and artificial intelligence in order to detect and treat illness in its earliest stages (9). Extraction of features using an existing system while building a machine learning classifier based atop, or refinement of the existing trained model with learned weights as starting parameters, are two examples of how transfer learning may be used in practise. Using a pre-trained CNN as a fixed feature extractor for the job at hand is typical practise when working with convolutional neural networks. In order to get the most out of transfer learning, it is not necessary to create new machine learning models for each and every opted activity. Healthy and cancer-affected lungs are shown in Figure 2.

In today's world, Transfer Learning as well as Artificial Intelligence are playing a crucial role, especially in the medical sector's diverse area. In order to identify the ailment, these computational intelligence tools do not hurt the patient (10). According to recent data, one of the most taxing duties in medical treatment is the diagnosis and categorization of lung cancer disease. In order to deploy algorithms for cancer classification and early detection, it is challenging to build a medical image dataset for training a machine learning model. For

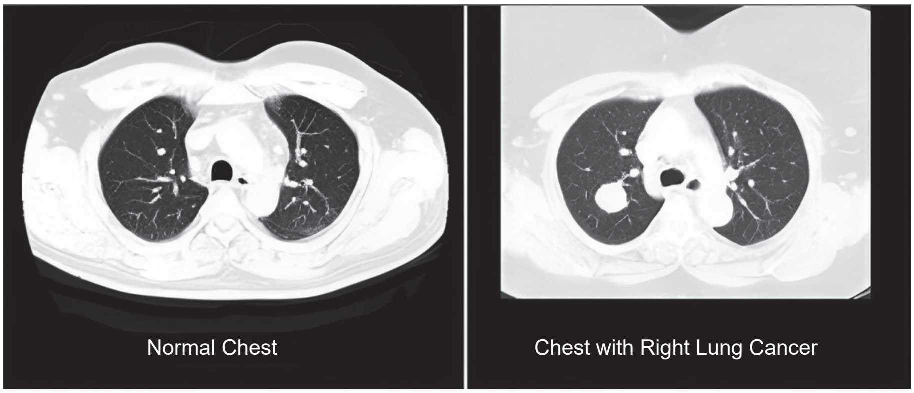

Fig. 2. Healthy and cancer-affected lungs.

example, we have considered a medical image dataset with 4,500 colour images divided into five categories, each of which had 900 images of cancerous or benign tissue in the colon or lung, as well as images of cancerous or benign squamous cell carcinoma in the lung (11).

The key contributions of this research articles are as follows:

- For the identification of lung cancer in CT scan images, this research offers an all-inclusive and relative scheme for assessing and relating computerised comprehensive and limited feature-extraction algorithms.
- A total of eight distinct kinds of pre-processed image characteristics are evaluated in this framework based upon the global and local feature modelling approaches.
- Three well-known machine learning methods, namely Support Vector Machines, K-Nearest Neighbors (KNN), and Naïve Bayes are also examined and compared with the proposed classifier model.

Literature survey

Researchers have used a variety of ways to analyse images, extract characteristics, and use machine learning techniques in order to diagnose lung cancer. Images of CT scans were segmented using image segmentation and subsequently classified using convolutional neural networks to detect lung nodules, according to Jin et al. Their method has a success rate of 85%. The researchers used a genetic algorithm to design a CAD system for detecting lung nodules. Depending on the intensity pixel value, they first segmented CT scans to identify areas of interest (ROIs) (12). To examine the images in all directions, they used varied thresholds. They employed an evolutionary algorithm to categorise each nodule after lowering the number of ROIs (depending on superior and inferior slices).

An overall sensitivity of 93.4% was recorded across 276 CT scans. It was discovered that the model could identify lung cancer based upon analysing the CT images by utilising textural cues and morphological procedures. The images were first clustered using the super pixel technique, and then morphological processes were applied. This contour -based algorithm was then utilised to identify the tumours in the images captured by the camera. Dice similarity was 84.88% for their system (13). The authors came up with a brand-new strategy for spotting lung nodules. With respect to force, form, quality, and framework, they used 128 characteristics. They were only able to achieve the sensitivity of 80%.

A KNN (K-Nearest Neighbour) classifier built on the nearest mean were used to assess SVM's performance. SVM outperformed the other approaches, according to their research (14). There is now a CAD method to categorise lung tumours into benign and malignant, according to Gonzalez and Ponomaryvo. Pre – processing, lung delineation, nodule identification, and classification were all incorporated in the proposed system. They used thresholding and morphological processes in the pre-processing stage to calculate multiple masks. Previous knowledge along with the similarity measure characteristics such as area, irregularity, concentricity, and partial measurement are characteristics was utilised to establish the Region of Interest (ROI). In this approach, they used the SVM technique to classify the data set. They recorded a 78.08% accuracy rate (15).

A novel CAD system was created by the authors. From the LIDC-IDRI database, they chose 420 instances at random. When looking for potential nodules, the system employed the Watershed approach to help separate them from other probable structures. Features were extracted using a HOG (Histogram of Oriented Gradients) method. SVM and a rule-based classifier were employed to minimise false positives. With a sensitivity of 93.9%, they used a cross-validation approach that has ten folds (16).

In order to detect lung nodules, a CAD system was created by Silva et al. In total, 333 tests were classified using SVM. The accuracy of their proposed approach was 95.21%. RSM (Random Subspace Method) was utilised by the authors to create a CAD system (17). RSM was used to identify pulmonary nodules in two stages of a supervised learning system. In the dataset of 126 samples, they collected 218 characteristics and used RSM and genetic algorithms to build a classifier. The proposed method has an accuracy rate of 88.9%.

To detect and identify lung cancers, the researchers employed EK-Mean clustering. They first used a median filter to eliminate the background noise. After that, they clustered and segmented the data using the K-means technique. The GLCM (grey-level co-occurrence matrix) was then utilised to excerpt characteristics like homogeneity, peak signal to noise ratio, correlation and entropy. That method has a 90.7% success rate. Linear Discriminant Analysis (LDA) was utilised by Aggarwal et al to categorise lung nodules and differentiate them from normal architecture. Segmentation was done using thresholding and grey-level features (18).

Using this method, an accuracy of just 84% could be achieved. Lung cancer was detected using CT scan images and marker-controlled watershed segmentation. Gabor filters were utilised as a pre-processing step to improve the image quality. An accuracy rate of 90% was reached in their method (19). In the segmentation process, an adaptive threshold technique was implemented. In order to detect lung nodules, the researchers followed a three- step procedure. A thresholding technique was first used to isolate the lung area in CT scans. Secondly, another active contour model was used to eliminate the lung vessels (ACM). They then used a form filter to identify the nodules. Finally, they employed a classifier to discriminate between real and false positive nodules based on their characteristics. The detection rate of this system was 85% (20).

A thresholding and region-growing technique was employed by Liu and colleagues. For segmentation, they employed pulmonary parenchyma, and for ROI extraction, they used a circle shape descriptor. The system's sensitivity was 85.6%, while its false positive rate was 13.4%. Solitary pulmonary nodules can be detected by authors (12), who created a computer-aided detection method. Segmentation was done using the best possible thresholding and neighbourhood (21). Dot enhancement filtering and angle histograms were employed to identify the nodule. Classification was accomplished by the application of support vector machines (SVM). A 97.5% sensitivity was achieved by the system. However, there were no further performance metrics available.

In the literature, there are a variety of approaches for extracting features. Grayscale contrast, grayscale correlation grayscale energy monochrome uniformity, sample variance, and Haar wavelet are among the possible candidates for an intensity histogram. Other possibilities include histograms of directed gradients and Gabor filters. For a variety of datasets, deep learning algorithms have been used to classify histopathology images of various cancers. For the pre-training of a neural network for the detection of lung and colon cancer, Garg et al in 2020 employed the LC 25000 image dataset. Colon and lung cancers were studied individually, and eight CNN-based algorithms were used to categorize cancer and normal images accurately. For the whole lung and colon dataset, there are no results for accuracy, F1 score, and other metrics (22).

Classification of lung cancer from cytological images using Deep Convolutional Neural Networks (DCNNs) achieved a 71% accuracy rate in 2017. Only 71% of classifications are accurate due to the implementation of DCNN, which consists of 3 conv layers, 3 pooling layers, along with 2 completely linked layers. Only lung cancer was considered when creating the image dataset. By cropping and resampling, the images were scaled to 256 x 256, which caused quantization errors. Deep learning algorithms and histopathological imaging can considerably benefit patients with lung cancer, according to Wang et al in 2019. After scaling the images to 300 x 300 pixels, researchers utilised a neural network to categorise images of healthy and cancerous tissues, but were only able to reach 89.8% accuracy.

Sophisticated sampling of tissue areas was used by Shapcott and colleagues in 2019 to improve performance and accuracy in colon cancer cell detection using deep learning. Deep learning was used to extract cell density and morphological parameters for colon cancer diagnosis and prediction, which the authors used for categorization. In 2021, Nur Ibrahim et al introduced a non-complex deep learning model that can detect four different types of colon cancer with an 83% accuracy rate for the dataset 2500, which includes tumour, complicated lymphoma, and stroma. Using histology images, the authors were able to extract characteristics from 150 x 50-pixel textures and present the results for four different forms of colon cancer. Smaller datasets have been utilised and the accuracy has also been reduced. Neither the sensitivity nor the selectivity of the classifiers can be determined (23).

Using Machine Learning, Wang et al in 2021 developed gender-specific lung cancer classification algorithms with 82.9% and 73.2% accuracy rates for women and men, respectively. It is necessary to do more research in order to identify any gender-specific characteristics which might or might not exist in order to improve accuracy. There is certainly room for improvement in this area, and we need a system that is more accurate regardless of gender. Deep residual learning was used by Bhatia et al to demonstrate a method for detecting lung cancer from a CT image. Preprocessing approaches have been described by the researchers as a pipeline for highlighting lung malignant spots and extracting the characteristics utilising models such as UNet and ResNet.

Many classifiers, such as XGBoost as well as Random Forest, are applied to the feature set, and the unique predictions are then combined to form ensembles for predicting whether or not a CT scan depicts malignant tissue (24). LIDC-IRDI outperforms traditional approaches in terms of accuracy by 84%. Extraction of FCH histogram features. Because of lighting fluctuations and quantization issues, it was also employed to minimise the noise, The purpose of these feature-extraction algorithms is image indexing and retrieval, not classification.

Proposed system

Preliminaries

The most important ideas and phrases in this literature are discussed in detail in this chapter.

Input images

Medical images come in a variety of forms. Lung cancer is frequently detected by the use of CT scan images. Accurate cancer detection systems require a comprehensive data set that includes images of both healthy and pathologic tissues in order to be successful. As a result, the computer is better able to distinguish between regular and aberrant images. One type of CT scanner captures each CT scan image. As a result, the coordinate system of each scanner was used to generate a range of distinct sizes and orientations. CT scan findings must be aligned with a structural model that replaces the initial system that was exclusive to every scanner for uniformity. Anatomical characteristics in each image are aligned to a shared-reference physiological coordinate system through image warping. Depending on how each CT scan image is obtained, certain undesirable regions such as areas outside the body, may be present. Image cropping is designed to eliminate all undesired areas from image graphs.

Feature extraction

Pixel intensity, brightness, histogram of directed gradients, Gabor and entropy filters as well as the grayscale sharpness and grayscale relationship and energy measures, as well as standard deviation and Haar wavelet, are a few of the possible contenders. For example, the intensity histogram is a visual characteristic that illustrates how frequently brightness of every pixel recurs. Thus, the number of intensity value repeats determines this characteristic, without regard to the image's pixel positions. There are several more features that may be calculated as follows: the histogram of directed gradients cells has been created within the image. The gradient's direction in each cell is then calculated. In the end, a histogram is constructed from the counts of each gradient direction. A Gabor filter consisting of a Gaussian window is used to examine the local surface frequency of the image. To further enhance an image's unpredictability and entropy is another characteristic that measures the scientific measurement of the unpredictability of pixel values. The entropy value (I_{en}) can be estimated by considering n normalised histogram numbers as:

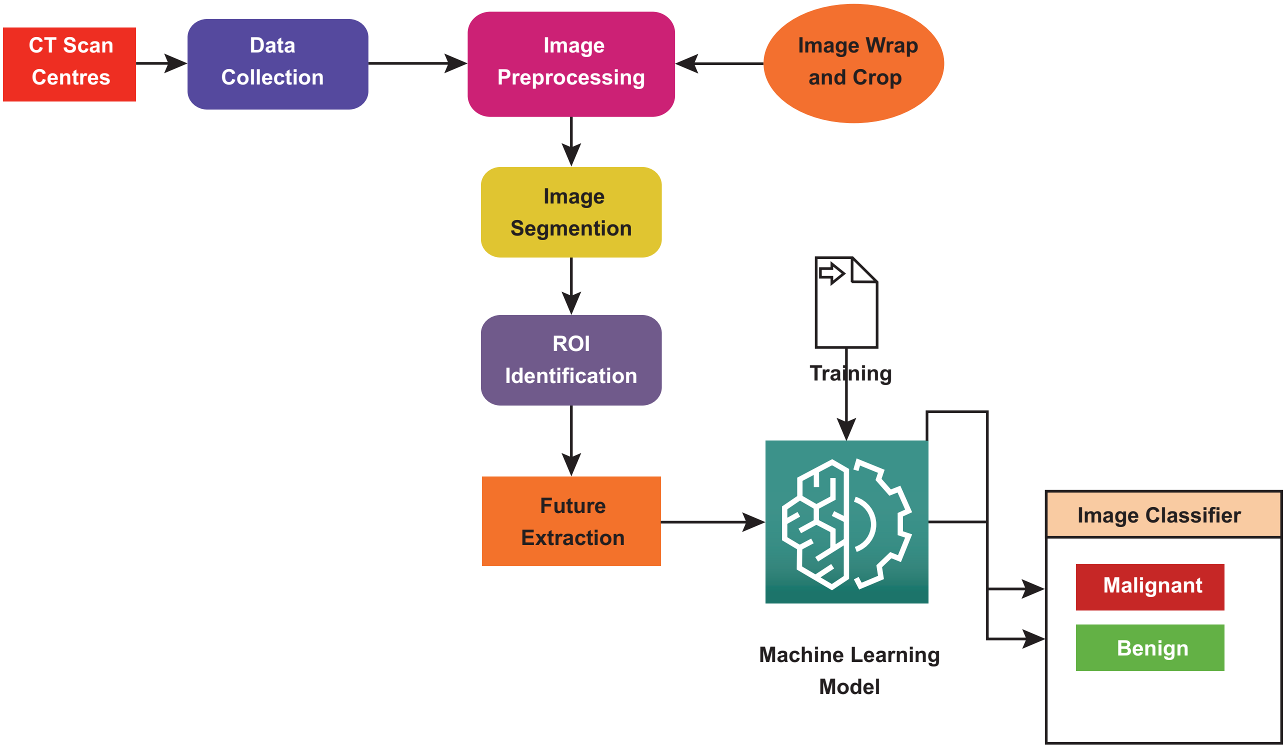

Fig. 3. Proposed system architecture.

$$I_{en} = -\sum n \log_2 n \tag{1}$$

All the grayscale features are providing a unique pixel value to define the image that is being considered. The contrast of the medical image (I_{con}) that gets the pixel differences of adjacent pixels of an image can be estimated as:

$$I_{con} = \sum_{x}^{Mh} \sum_{y}^{Mh} (x-y)^2 . n(x-y) \tag{2}$$

The adjacent pixel correlation ($Icor$) can be estimated as:

$$I_{cor} = \sum \sum \frac{(x,)-\mu_{\mathrm{r}}\mu_c}{\mu_{\mathrm{r}}\mu_c} \tag{3}$$

where μ_r *and* μ_c represent the mean pixel values of the row-wise and column-wise pixels, respectively. The diversity of pixels and their organisation pattern are analysed by the metric known as uniformity (I_e) that can be calculates as:

$$I_e = \sum_{x,} n\,(x,y)^2 \tag{4}$$

Gray-level co-occurrence matrix proximity to orthogonal is quantified using this metric known as homogeneity of the neighbourhood pixels I_h and can be calculated as:

$$I_h = \sum_{x,} \frac{n(x,y)}{1+|x+y|} \tag{5}$$

To find out the distribution of pixel data throughout the image, we are using the variance and the standard deviation measures of the given CT image. We are estimating the metric (I_{sd}) for P number of total pixels as follows:

$$I_{sd} = \sqrt{\frac{\sum_{x,}(n(x,y)-\mu)^2}{P-1}} \tag{6}$$

Also, we are assuming one dimensional Haar wavelet model for the discretised image as follows:

$$(l) = \begin{cases} 1, where\ 0 \leq l \leq 0.5 \\ -1, where\ 0.5 \leq l \leq 1 \\ 0, Otherwise \end{cases} \tag{7}$$

We can also consider two-dimensional images and the model can be represented as:

$$\omega^{x}_{y,a,b}(\mathrm{i},\mathrm{j}) = 2^{y/2}(2^{y}{}_{\mathrm{i}-a}, 2^{x}{}_{\mathrm{j}-b}) \tag{8}$$

Machine learning models

A technique must be devised to identify normal as well as abnormal sections in the images after characteristics have been retrieved. There are numerous machine learning algorithms available. Random forests, SVM, decision trees, KNN (K-Nearest Neighbours), and Naïve Bayes are well-known approaches. Classifying the feature vectors into abnormal and normal categories is done using the Support Vector Machine (SVM) approach. Analysis of a labelled collection of vectors is used to learn classifications, which are subsequently used to categorise unlabelled vectors as belonging to one of two distinct classes. This classifier determines the optimum hyper-plane for dividing the subspace into two distinct areas, then uses SVM to distinguish the two classes. The proposed system architecture is shown in Figure 3.

With this hyperplane, it is possible to maximize the margin, ensuring that feature vectors closest to the hyperplane on either side have the most space within the partitions. It is the decision boundary, a hyper-plane that divides the different classes of people. Class 1 is reserved for data points above this line. Linear SVM is the term used to describe the approach that calculates the values of and b with the widest margin available. Each extracted feature is evaluated to a set of previously labelled feature vectors. The majority of adjacent labelled feature vectors are used to derive the label for the feature vector. In order to discriminate between positive and negative examples, the Random Forest learning approach uses a series of heterogeneous decision tree models. A random sample of a dataset is used to build each decision tree. Decision tree models are used to compare feature vectors to the previously generated decision tree models. All of the decision trees' predictions are gathered together. The feature vector's label is chosen by acclamation from all of the decision tree predictions that have been made.

The cross-validation approach may be used to test and evaluate the results of various feature extraction techniques and deep learning systems. All labelled feature vector sets are separated into eight separate parts, with seven of them utilised in training and one in testing, respectively. There are 8 repetitions of the method to ensure that each division is tested once. Each performance metric is given ten values, which are then aggregated. There are a variety of ways to measure performance. Within research, some researchers have employed measures of efficiency, utilising specificity as along with sensitivity to measure precision. It is the system's capacity to discriminate between normal and abnormal instances that is the measure of accuracy. Specificity demonstrates the system's capacity to recognise normal cases, while sensitivity demonstrates the system's ability to identify aberrant cases.

In order to develop global models, visual characteristics are taken from the complete image. As a result of extracting these various forms of information, a feature vector is created. Using them, the global recognition models may be computed and constructed. The samples have been twisted, gathered, followed separated among a number of local chunks in order to be able to locate the image areas containing questionable material. The image characteristics of the local block are extracted in each local block. Local blocks are represented by a feature vector that includes the extracted feature types. They are then utilised to create and calculate the local detection methods.

Algorithm 1: Feature Extraction
Input: Set of CT images $I_1, I_2, I_3 \ldots$
Output: Features from the images

```
1.  For each image $I_n$ do
    1.1  Find co-ordinate values of image
    1.2  Mark anatomical control points of image
    1.3  Complete Image warp process
2.  Compute geometric transformation for the sample image
3.  Crop the image by choosing cropping points as $I(x, y)$ such
    that x and y are extreme points of the image
4.  Generate normalised intensity histogram by estimating
      $I_{en} = -\sum n \log_2 n$.
5.  Calculate energy of the image as $I_e = \sum_{x,y} n\,(x, y)^2$
6.  Find the contrast value with $I_{con} = \sum_{x}^{Mh} \sum_{y}^{Mh} (x - y)^2 . n(x - y)$

    1. Establish the correlation matrix by calculating correlation values of the pixels from.

       $I_{cor} = \sum \sum \frac{n(x,y) - \mu_r \mu_c}{\mu_r \mu_c}$.

    2. Extract the standard deviation feature with

       $I_{sd} = \sqrt{\frac{\sum_{x,y} (n(x,y) - \mu)^2}{P-1}}$

    3. Establish Haar wavelet model as
       $m^{x}_{y,a,b}(i, j) = 2^{y/2}(2^{y}_{i-a}, 2^{x}_{j-b})$.
```

To create the final lung cancer prediction model, this study evaluates the effectiveness of each feature extraction and learning approach. Each image in our dataset is subjected to two image preprocessing steps: image warping and image cropping. Anatomical control points were manually indicated on 19 images to ensure that the structural characteristics were aligned using a consistent structural organisation of the model. The corners of the lungs and the corners of the heart were the sites of these control points. The images were then warped to match with the respective chokepoints in our standard anatomical reference frame using these 19 control points.

Proposed methodology

To develop a lung cancer diagnosis system using CT scan images, we present a framework for extracting global and local features that is both comprehensive and comparable. Features gleaned from the images are subjected to a bevy of categorization algorithms. A Data Collection Process, the Strategic Process, and the Local Process are the three key sequential phases of this technique. The samples of medical images are gathered throughout the Data Collection stage.

The acquired images are normalised with image distortion and clipping during the Global Phase. Each pre-processed image is then used to extract 8 distinct kinds of global characteristics. Each image receives 8 feature vectors as a result of this process. Machine learning methods are then applied to these feature vectors to create detection models. Afterwards, each detection model's performance is tested and compared to that of the other detection models.

The images are partitioned into a number of regional blocks that construct ROIs during the Local Phase. Feature vectors are generated from the extracted feature of the image blocks using the kinds of visual features that worked well all through the Strategic Phase. Using the learning techniques that worked effectively in the Strategic Phase, those feature vectors are therefore utilised to create machine learning models. To find the best number of blocks for each image, we test with varying quantities of blocks per image, including just one block. The performance has been evaluated and related to the performance of the existing approaches.

This study relied on a collection of 800 CT scans. In total, there were 400 unusual instances and 400 normal instances. The images were chosen at random from Kaggle's collection of thousands of images. Prior to undergoing any therapy or surgery, these individuals underwent diagnostic CT scans. A CT scan produces a stack of slices, and each image reflects a particular slice from that stack. A CT scanner's slice thickness ranges between 3 and 6 mm. There are more than 3,000 images in this collection from a variety of patients. The training phase of our algorithm to determine lung cancer using CT scan data in the global phase. Data pre-processing, extraction of features, and feature learning are all steps in the process. In the testing stage, the model's performance is evaluated based on the results of the learning process. Various feature extraction and learning techniques are applied to the respective chokepoints in our standard anatomical reference frame using these 19 control points.

An anatomical coordinate system reference image and an image to be warped were marked with control points using a MATLAB programme. The control points in the chosen image were then aligned with those in our standard anatomical reference frame using a geometric transformation. For each of the 19 coordinates in the sample image, an non-linear conversion was calculated that alters the selected one to match its (x,y) pixel coordinates that these already in the source images.

To create distorted copies of all 1000 images in our dataset, we used a geometric modification that we calculated and applied to each individual image. As with the originals, these images have the same level of textural detail, but the control points have been aligned to match those in the source images instead. The previously twisted image is clipped. The topmost and bottom most coordinates, as well as the maximum and minimum points, were the cropping locations. Our suggested methodology relies heavily on the characteristics extracted from extracted images. Various feature extraction approaches are compared in this research. These feature categories are extracted from the entire image for global feature extraction.

Extracted features for each image are generated using this method. After that, the learning process makes use of these feature vectors. To represent each image, a feature vector is constructed using the retrieved features. In order to develop the final detection model, these 8 feature vectors are employed as inputs in the learning process. To do this, a variety of algorithms are employed. The very similar pre-processing and segmentation techniques procedures were used for every CT scan image. The learned algorithm categorises medical images as either benign or malignant. Table 1 lists the feature-extraction process steps and identified features of the proposed system.

```
Algorithm 2: Classification Model
Input: CT image from the training and testing dataset represented with $I_1$, $I_2$, $I_3$ … Output: Classified data as benign or malignant
Steps:
1. Get the training data images for training the classifier model.
2. For each image $I_n$ do
     • Get the features by feature extraction model.
     • Estimate the correlation value as $I_{cor} = \sum\sum \frac{n(x,y)-\mu_r\mu_c}{}$
3. Estimate the homogeneity from the equation
     $I_h = \sum_{x,y}^{\mu_r\mu_c} \frac{n(x,y)}{1+|x+y|}$
4. Compare with the discretised image model
     $m(l) = \begin{cases} 1, where\ 0 \le l \le 0.5 \\ -1, where\ 0.5 \le l \le 1 \\ 0, Otherwise \end{cases}$
5. Find the $(l)$ value for the testing image.
6. If $(m(l))$ for image $I_n = 1$ then
     Conclude benign Else
     Conclude malignant
7. Compute loss function value for the classifier model.
8. Update the learning function value to reduce the loss value.
9. Obtain the metric values to analyse the performance of the classifier.
```

The efficiency, specificity, and precision of each of the 60 learned models were calculated using a k-fold cross-validation approach. Global Training and Testing functions compare three machine learning algorithms with extracted image data to see

Table 1. Feature extraction and identified features of proposed system.

Name of the feature	Estimator
Entropy value (I_{en})	$-\sum n \log_2 n$
Contrast of the medical image (I_{con})	$\sum_x^{M_h}\sum_y^{M_h}(x-y)^2 . n(x-y)$
The adjacent pixel correlation (I_{cor})	$\sum\sum\frac{(x,y)-\mu_r\mu_c}{\mu_r\mu_c}$
Uniformity (I_e)	$\sum_{x,y} n\,(x,y)^2$
Homogeneity (I_h)	$\sum_{x,}\frac{(x,y)}{1+\lvert x+y\rvert}$
Standard deviation (I_{sd})	$\sqrt{\frac{\sum_{x,}(n(x,y)-\mu)^2}{P-1}}$
One dimensional Haar wavelet model	$(l) = \begin{cases} 1, where\ 0 \le l \le 0.5 \\ -1, where\ 0.5 \le l \le 1 \\ 0, Otherwise \end{cases}$
Two-dimensional Haar wavelet model	$\omega^x_{y,a,b}(\mathrm{i},\mathrm{j}) = 2^{y/2}(2^y_{\mathrm{i}-a}, 2^x_{\mathrm{j}-b})$

Table 2. Prediction Accuracy Comparison.

Classifier	Prediction Accuracy (%)
SVM	84.8
KNN	86.5
Naïve Bayes	83.67
Proposed	87.52

which techniques outperform the others in terms of performance. Local training and testing functions then make advantage of the more effective features and techniques. Both global and local images warp and crop in the local phase of this process. Blocks are created for each twisted image in this step. We then employ the sorts of feature extraction that worked well during the Strategic Stage in each of the individual blocks.

With *P* being the maximum number of pixels in the image, every feature type generates b feature vectors per image. Afterwards, each of the *P* feature vectors may be utilised to label a single block. All 400 anomalous images are examined to see which blocks have suspicious material in order to cut down on training and learning time. As a result, the feature extraction and training and learning algorithms do not include many blocks in anomalous images that do not contain any worrisome material. To put it another way, we only train our algorithms using the blocks of questionable material that appear in the anomalous images. Every ROI (Region of Interest) in a test image is labelled as malignant or benign using these models in the test procedure.

If an aberrant ROI is found in a test image, the entire image is considered abnormal. No abnormalities are found in any of the images; hence the image is categorised as normal. It is good to have training sets that comprise an equal amount of aberrant and normal examples for the purpose of training. That is why we went with 400 scans of pathological tissue and 400 scans of healthy tissue. In contrast, when the images have been broken down into individual blocks, there are numerous standard chunks compared to aberrant blocks. In each pathological tissue image, there may only be ten abnormal blocks in the 400 aberrant images. As a result, the number of normal blocks vs anomalous blocks inside the training set with that block is drastically skewed.

Results and discussion

On a system with an Intel Quad Core CPU at 3.7 GHz and 16GB of RAM, MATLAB is used to conduct experiments. As previously stated, lung nodules from the Kaggle dataset are evaluated. Here, an image database is randomly partitioned into different sets via a validation mechanism. Standard measures of specificity, sensitivity, and accuracy are used to compare the performance of different systems. Only the most relevant traits are considered, and they may not be quantified for categorization purposes. Table 2 compares various classifier algorithms and their prediction accuracy.

Methods for selecting the most important distinguishing characteristics are anticipated. A 10-fold cross validation and classification are then done using the Improved Nave Bayes classifier. Predicted feature selection approaches are evaluated using classification accuracy and the strategy is compared with the most commonly used feature selection methods. A comparison is made between the performance of the upgraded Naïve Bayes classifier and the support vector machine (SVM) and k-NN classifier.

The expected I-NB classification is investigated using the classification rate, which is the ratio of properly classified images to entirely classified CT lung images, specificity (SP), and sensitivity (SN). They performed better in the general strategy, the precision, responsiveness, and selectivity rates of the local feature extraction approach are described and analysed a 1 x 1 block to a 20 x 20 blocks. For eight image characteristics, SVM and generic model assesses the correctness, consideration and precision. Comparative analysis of prediction accuracy is depicted in Figure 4.

When utilising SVM with global feature extraction, the findings show that outcomes are obtained. A total of 78%, 79%, and 81% of the metrics of performance were attained by the Gabor Filter, making it the top performer. However, we initially carried out multiple tests to discover the optimal k values for the global feature extraction technique of accuracy. Using a universal feature extraction technique, it reveals the k values that produced the greatest accuracy rates. For the 8 image features, KNN with global feature extraction was used to test accuracy, sensitivity, and specificity. Confusion Matrix for the proposed system is shown in Figure 5.

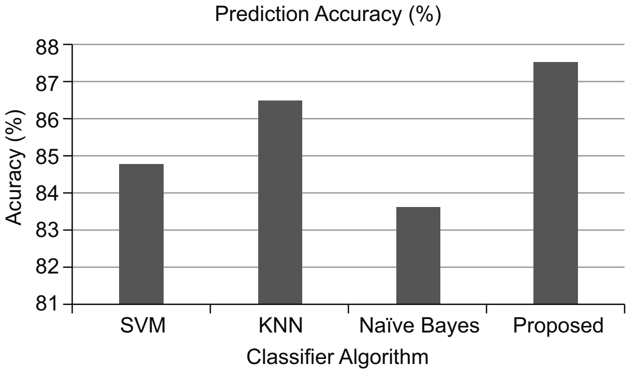

Fig. 4. Comparative analysis of prediction accuracy.

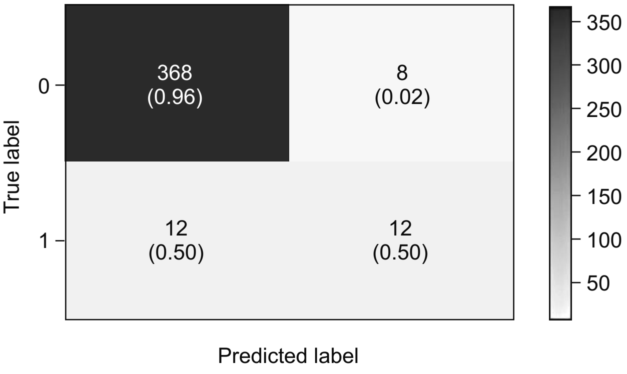

Fig. 5. Confusion matrix for the proposed system.

Table 3. Performance comparison of classifiers.

Classifier	Precision	Recall	F1 Score
SVM	88.92	89.56	0.854
KNN	85.65	88.56	0.845
Naïve Bayes	87.65	87.52	0.889
Proposed	91.05	90.12	0.915

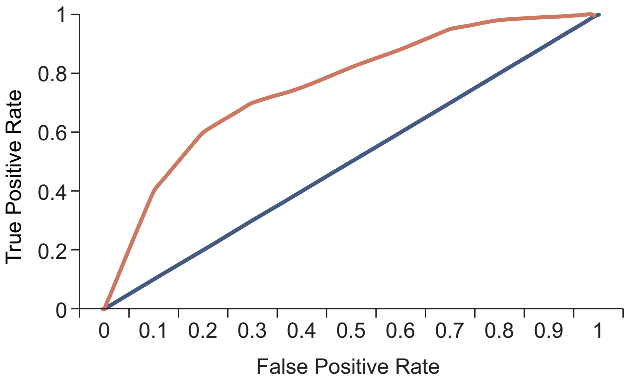

Fig. 6. ROC curve.

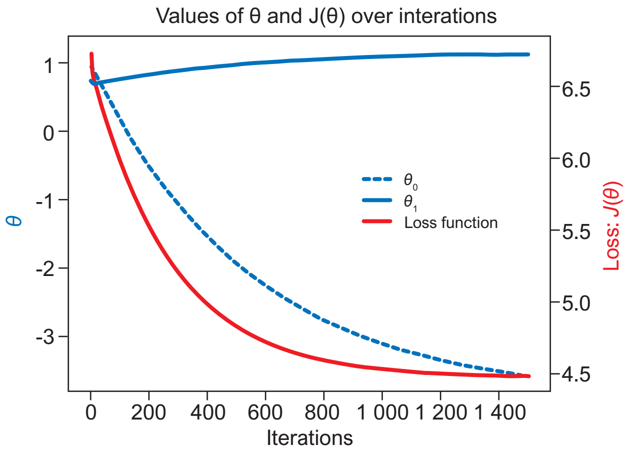

Fig. 7. Training of the model.

In terms of overall performance, the Gabor Filter was the most accurate with an accuracy of 65%, a sensitivity rate of 61%, and a specificity rate of 70%. A Nave Bayes method employing global feature extraction produced better results with the others. Its overall accuracy, sensitivity and specificity were all found to be at 66%, 60% and 71%, respectively, for the Gabor filter. ROC curve of the proposed system is shown in Figure 6.

The findings also show that the Gabor Filter outscored the other feature extraction when using the proposed technique. We use a dataset of 800 medical images and the results of this study show that such local feature extraction strategy is superior to the global approach. In addition, it implies that the suggested strategy outperforms the existing approaches described within this study. For a variety of reasons, the study described here is superior to others that have been conducted in the past. Performance comparison of classifiers is shown in Table 3.

As a first step, the study used a far bigger data set than is commonly used in the scientific literature. Compared to the literature, our study relies on a much larger data set of 800 CT scan images. Secondly, our study outperformed the approaches listed in the literature in terms of performance. An efficiency, sensitivity, and precision rates of 97% were reached. Also, whereas the mentioned study employed global detection approaches utilizing diverse machine learning algorithms, our research focuses on building the localized machine learning approach for detecting lung cancer. Figure 7 explains the training process of the proposed classifier model.

Anatomical areas in CT scan images are critical to the interpretation and diagnosis by radiologists, hence the building on localized learning models is essential. It's possible that normal material in one section of the body is abnormal in another. We also compared the efficiency of 8 distinct characteristics that were extracted and three existing approaches in our study. This explains why our suggested local segmentation and learning technique is superior to other methods.

Conclusion

This study provided a methodology of global and local extraction of features for the diagnosis of lung cancer in CT scan images that is both comprehensive and comparable. In this work, three existing machine learning algorithms were evaluated using distinct image feature extraction approaches using internal and external feature extraction strategies. Anatomical structure, local extracted features, and model learning were enabled through image warping. There were eight different features extracted and tested in this study. The results showed that the proposed approach to feature extraction techniques surpassed all the existing approaches. The proposed feature extraction surpassed the typical global technique over study outcomes. Results from this study reveal that the suggested strategy outperforms other referenced methods within the research not only in terms of obtaining greater accuracy but also in terms of utilizing 800 CT scan data and developing localized teaching methods for lung cancer diagnosis. Radiologists may be able to better diagnose lung cancer by utilizing SVM combined with Gabor Filter extraction of features to identify worrisome spots in CT scan images.

References

1. **Abdullah ANS.** A review of most recent lung cancer detection techniques using machine learning. Int J Sci Bus 2012; 5 (3): 159–173.

2. **ALzubi B, Kannan B, Tanwar S, Manikandan R, Khanna A, Thaventhiran C.** Boosted neural network ensemble classification for lung cancer disease diagnosis. Appl Soft Comput 2019; 80: 579–591.

3. **Sanmartin AP, Jabba D, Jimeno M.** Applications based on service-oriented architecture (SOA) in the field ofhome healthcare. Sensors 2017; 17 (8): 1–16.

4. **Baratloo MH, Negida A, El Ashal G.** Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Arch Acad Emergency Med 2015; 3 (2): 48–49.

5. **Benharref MA, Serhani.** Novel cloud and SOA-based framework for Ehealth monitoring using wireless biosensors. J Biomed Health Inf 2014; 18 (1): 46–55.

6. **Clark B, Vendt K, Smith J, Freymann J, Kirby P, Koppel S, Moore S, Phillips D, Maffitt M, Pringle L, Tarbox, Prior FD.** The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J Digit Imag 2013; 26 (6): 1045–1057.

7. **Demirkan DD.** Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decis Support Syst 2013; 55: 412–421.

8. **Forkan I, Khalil A, Ibaida Z, Bdcam T.** Big data for contextaware monitoring – A personalized knowledge discovery framework for as sisted healthcare. IEEE Trans Cloud Comput 2017; 5 (4): 628–641.

9. **Gazzarata F, Vergari T, Cinotti S, Giacomini M.** A standardized SOA for clinical data interchange in a cardiac telemonitoring environment. J Biomed Health Inf 2014; 18 (6): 1764–1774.

10. **He X, Fan YL.** Toward ubiquitous healthcare services with a novel efficient cloud platform. Trans Biomed Eng 2013; 60 (1): 230–234.

11. **Hsu CH, Hidayati SC, Cheng WH, Chen YJ.** Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Ther 2022; 8: 2022.

12. **Carvalho IR, Tripathy AK.** Knowledge discovery in medical systems using differential diagnosis, LAMSTAR, and K-NN. Trans Inf Technol Biomed 2012; 16 (6): 1287–1295.

13. **Zhang JYC, Jin QL.** Pulmonary nodule detection based on CT images using convolution neural network. Proc. 9th Int. Symp. Comput. Intell. Design (ISCID), Dec. 2016: 202–204.

14. **Garg KN, Kaur D.** Segmentation and feature extraction of lung region for the early detection of lung tumor. Int J Sci Res 2014; 3 (6): 2327– 2330.

15. **Kim HC.** Service-oriented architecture structure for healthcare systems utilizing vital signs. IET Commun 2012; 6 (18): 3238–3247.

16. **Kuo RF, Chang DR, Chen, Lee CC.** Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Res Treat 2001; 66 (1): 51–57.

17. **Lasierra A, Garcia AJ.** Designing an architecture for monitoring patients at home: Ontologies and web services for clinical and technical management integration. J Biomed Health Inf 2014; 18 (3): 896–906.

18. **Ma WW, Xia B, Zhang S, Yuan H, Jiang H, Meng W, Zheng X, Wang X.** Multiplexed serum biomarkers for the detection of lung cancer. E Bio Med 2016; 11 (9): 210–218.

Received September 22, 2023.
Accepted November 10, 2023.