International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 5, October 2023, pp. 5135~5146
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i5.pp5135-5146

Classification techniques using gray level co-occurrence matrix features for the detection of lung cancer using computed tomography imaging

Shankara Chikkalingaiah[1], Subbarao Anantha Padmanabha Rao Hari Prasad[1], Latha Dabbegatta Uggregowda[2]
[1]Department of Electronics and Communication, Faculty of Engineering and Technology, Jain University, Bengaluru, India
[2]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning),
Vidyavardhaka College of Engineering Mysuru, Karnataka, India

Article Info

Article history:

Received Aug 22, 2022
Revised Feb 4, 2023
Accepted Mar 9, 2023

Keywords:

Decision tree
Grey level co-occurrence matrix
Median filter
Random forest
Support vector machine
Thresholding

ABSTRACT

Lung cancer, which causes the majority of fatalities worldwide each year, is one of the deadliest diseases. The survival rate of cancer patients could be improved with better cancer detection methods. Image processing and machine learning have both been used to aid in lung cancer detection, but a method that both increase accuracy and increases a patient's survival rate has yet to be identified. In an effort to find the most effective method for the accurate lung cancer recognition, this paper analyses and compares several classification algorithms. Lung computed tomography (CT) images are enhanced by removing noise using a median filter. For filtered image, threshold segmentation is used to segment it into distinct parts. From the segmented image different features are extracted using the grey level co-occurrence matrix (GLCM). several classification strategies, including support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and decision tree (DT) methods, are used to classify lung images as malignant or normal based on the extracted features. Methods are evaluated based on a number of various performance measures, like accuracy, a precision, the recall, and the F1-Score. Based on the experimental outcomes, SVM outperforms other classification methods in accurately detecting lung cancer with an accuracy of 99.32%.

Corresponding Author:

Shankara Chikkalingaiah
Department of Electronics and Communication, Faculty of Engineering and Technology, Jain University
Jain Global Campus, Jakkasandra Post, Kanakapura, Main Road, Ramanagar, Bengaluru, Karnataka, India
Email: shankara151987@gmail.com

1. INTRODUCTION

Among the most typical types of cancer, the most serious disease is lung cancer in the modern world. Because of this, the number of people who die from it keeps going up. By 2030, 2.45 million people are expected to pass away from lung cancer, a 39% increase from the cases that were reported in 2018 [1]. Patients with lung cancer have a mortality rate equivalent to approximately 14% (across all stages), and their average duration of survival is only about 5 to 6 years. Lung cancer is more likely to develop among smokers, and as a result, lung cancer kills more men and women than any other disease because of tobacco products or indirect smoking [2]. The primary cause of lung cancer of the non-small cell type (NSLC), which makes up 85%, and small cell lung cancer (SCLC), which makes up 15%, is tobacco use (smoking or chewing) [3]. The size of the tumors as well as the degree to which cancer spreads throughout the lung determine the stage of the disease. Based on how severe it is, lung cancer is categorized into 4 stages [4].

With varied degrees of effectiveness, many methods of image processing and automated classification systems have been identified in numerous studies. Both traditional lung analysis and automated classification methods produce different outcomes. The standard method for detecting lung cancer includes several stages that might increase computation complexity and error rate when using image processing techniques. It has a negative impact on detection performance, computational efficiency, and classification error rates. The majority of the aforementioned techniques, however, are costly, time-consuming, and less efficient in detecting malignant tumors.

The important contribution made by the proposed research work lie in the precise detection of cancer in lung, which improves the survival rate and provides significant importance in solving societal issues, thereby contributing to the novelty. Usually, after a thorough physical examination of computed tomography (CT) scans, lung cancer is confirmed by physicians, which requires a lot of time and is not always accurate. The majority of these conventional and physical computational methods are very expensive, time-consuming, and less effective in detecting cancer tumors. Many investigations have evaluated various methods of image processing and automated classification systems with variable degrees of effectiveness [5]. Lung cancer, if detected accurately, reduces the risk of surgery, enables various treatment possibilities for recovery, and thereby increases the survival rate. The proposed research provides standard state-of-the-art optimization machine learning methods on feature extraction, segmentation, as well as classification to accurately identify the nodules in the lung CT scans for the recognition of cancer in the lung. The machine learning (ML) methods support vector machine (SVM), k-nearest neighbor (KNN) method, random forest, (RF) and the decision tree (DT) evaluated the CT image dataset for texture feature classification. The research work produces promising results for lung cancer detection by providing the best-suited algorithms for the detection and classification of lung tumors accurately.

The researcher used the thresholding method and segmentation using a marker-controlled watershed, and image smoothing using a Gaussian filter. The binarization method is applied to determine overall number of pixels that are black and white. The lung tumor is identified by setting a threshold value. If the number of black pixels is higher than the threshold value, then the tumor is considered to be normal. Consequently, if the number of black pixies is less than the cutoff point, the tumor is cancerous [6]. This technique has the drawback of producing less accurate results when the lung image and background pixels are significantly out of proportion. Using appropriate feature extraction techniques, the needed features are not extracted. It is difficult to figure out a global threshold that yields adequate results whenever the background of an image is unpredictable. Contrast limited adaptive histogram equalization (CLAHE) as well as fuzzy c-means (FCM) methods are applied for image enhancement and segmentation, respectively. The Bayesian Classifier, which is a form of probabilistic modelling is applied to determine whether the lung input image is malignant or not. Grey level co-occurrence matrix (GLCM) is used to extract the features [7]. With small datasets, the Bayesian classifier can be used, but it couldn’t be more perfect for situations involving complicated classification problems, and it can only be used when the features are different from one another. The SVM classifier and KNN algorithm were used to accurately diagnose lung cancer. The methodology is divided into four stages: preprocessing the data, segmentation of filtered image, feature extraction, and the classification. The discrete wavelet transformation method of segmentation is used to select the region of interest initially after the CT scan image has undergone pre-processing. The segmented image is applied to extract features from the GLCM, such as the correlation, entropy, variance, contrast, energy, and dissimilarity. Binarization is applied to determine the cases of lung cancer probability. SVM is used to classify the image and determine whether it is a malignant tumor or normal [8]. This proposed technique has the drawback of utilizing an advanced degree of the algorithm as well as the extreme gradient boosting algorithm for improved data set utilization.

Using a decision tree algorithm, the author developed a strategy for cancer diagnosis. The database contains the input images obtained from the UC irvine machine learning repository, and the data set is separated into training and testing datasets. If it can decide whether all cases belong to the same class, the tree is a leaf in the first phase, and that node is labelled by that class. Next, after calculating the entropy and information gain of each characteristic, the best selection criterion will be selected. On the basis of the information obtained, a decision tree is constructed and then trained with the help of the training dataset. The approach predicts lung cancers in the input CT scans when evaluated using a testing dataset [9]. The decision tree algorithm drawbacks include the model training process takes longer timeframe, is higher cost, and inadequacy for regression and continuous value detection. The CT images of the lungs were divided into cancer and normal lungs using a random forest method, according to the author CT image from the lung image database consortium (LIDC) standard dataset, images are filtered with median filter to remove noise and also applied Gaussian filter to smooth the image. Watershed segmentation follows preprocessing. Using the images that have been segmented, different features like area, the diameter, an eccentricity, the centroid, as well as mean intensity are extracted. These extracted features are given to classification model like random forest classifies images as cancerous or normal. The model's accuracy, sensitivity, and specificity values were respectively 89.90%, 90.85%, and 88.32% [10]. The random forest technique has the disadvantage of requiring more capacity for computation as well as resources since it creates numerous trees to integrate their results. The training procedure for the models requires more time since there are more trees to train, and it is also less interpretable.

In the preprocessing phase, to increase the quality of lung images, the Gabor filter is used, which results in accurate cancer identification in the lung. The lung input image was segmented using three different techniques like watershed segmentation, region growing, and watershed with marker control and masking. They are analyzed in terms of the amount of time required for processing, and segmentation masking is chosen since it takes considerably less time than other methods. In order to extract features for lung cancer analysis, the color attribute is utilized. The lung condition is then determined, like normal or malignant, by comparing black pixels with the threshold value of 17,178.48 [11]. Histogram equalization techniques and the conversion of red green and blue (RGB) to grayscale are utilized during the pre-processing stage. Images are segmented using the image thresholding technique, and also region props method is used to extract six features after it has been segmented. A network with forward and backward propagation is used in order to categorize the CT images of the lungs. It is constructed using MATLAB and employs a sigmoid training function at all layers, achieving 78% efficiency [12]. The proposed method’s disadvantage is that the model is trained with just 70 images. The model may be trained with a larger database, which will improve the method's performance. The CLAHE method is being considered for pre-processing. It improves the contrast among the input images of the lung to improve image quality and limit noise amplification. Images from the LIDC dataset were fed into the GLCM for feature extraction. This technique used the input image to gather texture characteristics like a contrast, correlation, an energy, sum, maximum likelihood, the variance, and dissimilarity. The characteristics that are extracted are spatial real value data, which is essential for categorization. The SVM classification algorithm was used to classify the training and test images using the extracted features, and obtained an accuracy of 79.166% [13].

The researcher used a median filter to lung CT images in grayscale that eliminates noise like salt and pepper and aids in the precise identification of lung cancer. In addition, it was proposed that a Gaussian filter could be used to smooth the input image and get rid of any speckle noise that was present. Using the watershed approach for segmentation enables the separation and identification of items that contact the image and further segments lung cancer nodules that are touching spurious nodules. Features including area, the perimeter, the diameter, an eccentricity, the centroid, and mean intensity are fed into the classifier using segmented images. Then, SVM method is applied to accurately classify images as malignant or normal with an accuracy of 86.6% [14]. To remove the artifacts that are added during the acquisition of input CT images, the author used a Wiener filter in the preprocessing step. The nodule structure is separated from the lung input image portion using the automated region growth (ARG) technique. Moreover, to remove irrelevant clusters, two-dimensional structure or size parameters like eccentricity and area are used. In order to remove blood vessels from each nodule in the subsequent slices, a centroid analysis with three-dimensional is performed to calculate the centroid change for every nodule. Three-dimensional texture characteristics are also extracted to eliminate calcifications. The SVM method is applied to perform classification on extracted features, which achieved 94% accuracy [15]. The researcher applied a method for removing noise using a median filter during the pre-processing stage. The super pixel density-based region (SPDBR) technique was utilized to segment the filtered images using internal structures. The SVM model is then applied to perform classification of lung nodules using characteristics. Using a nonlinear SVM approach, the lung nodule candidate locations were divided into the malignant nodule and normal nodule judgments, and 84.75% accuracy was attained [16].

To make accurate predictions about lung cancer, several researchers employed various classification techniques, various sample sizes, and various datasets. As a result, standardizing the classification algorithm and overcoming the existing constraints, lung cancer detection methods is essential. The aim of this research is to find the effectiveness of all various classification algorithms in accurately identifying cancer in lung and this also determining the optimum classification model in the proposed method.

The remaining part of this proposed research paper is summarized: the methodology is described in section 2. The results, discussions, and comparisons of the various classification techniques are explained in section 3. Section 4 of this research paper provides an explanation of the conclusions.

2. METHOD

The proposed methodology is conducted out on lung CT image datasets to accurately classify lung cancer. The four steps involved in lung cancer detection are pre-processing to remove noise, segmentation to get the lung segmented image, feature extraction to extract features, and classification method. The proposed technique begins with pre-processing CT images which are obtained out of the LIDC database to eliminate different kinds of noise that was produced while the image was being acquired. A lung region of interest is extracted using segmentation techniques from the filtered image. Texture features of lung CT images are extracted using a standard GLCM technique with various optimized novel features. Different kinds of feature extraction techniques are used to extract features of the segmented image, such as its energy, entropy, contrast, and homogeneity. This optimized resultant feature subset is then trained by various machine learning models to accurately predict and detect the type of lung cancer and compare it with other algorithms to determine the best performance model. In order to accurately detect and predict lung tumors, the researcher proposed a hybrid framework with various optimized machine-learning algorithms. In addition, the survival rate can be improved with the proposed method. Different performance metrics are utilised in order to conduct an effectiveness analysis of different algorithms on input CT images of the lungs in order to select the most appropriate algorithm for tumor detection. In the proposed method, four different non-parametric ML classifiers like SVM, RF, KNN and the DT methods are applied to a feature dataset to accurately classify them as either cancerous or benign. As shown in Figure 1, the best classification model is determined by comparing the different techniques with respect to their performance parameters, like accuracy, recall, F1 score, and precision.

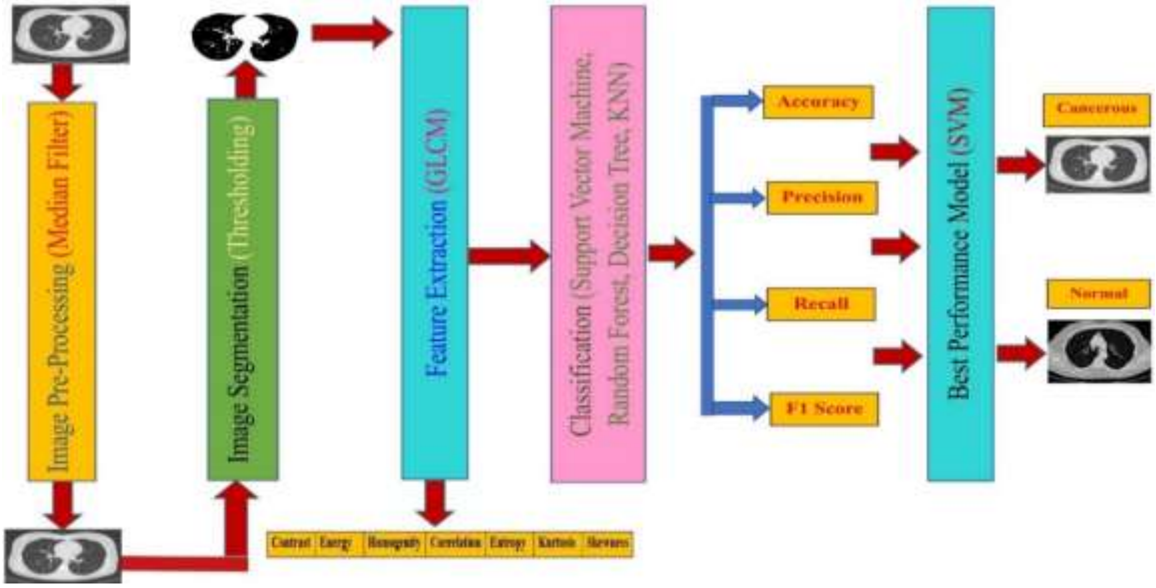


Image Pre-Processing (Median Filter)
Image Segmentation (Thresholding)
Feature Extraction (GLCM)
Classification (Support Vector Machine, Random Forest, Decision Tree, KNN)
Accuracy
Precision
Recall
F1 Score
Best Performance Model (SVM)
Cancerous
Normal
Contrast
Energy
Homogeneity
Correlation
Entropy
Kurtosis
Skewness

Figure 1. Proposed methodology

2.1. Dataset

The very first method of medical image analysis is the acquisition of images. The automated proposed method utilized a collection of images as input and used them to classify nodules in the lung. The LIDC dataset, which contains scans of 1,018 unique individuals, selected as the source for the input CT images. The CT scans of each individual patient contain from around 150 to 550 images in the DICOM format that are utilized to analyze and compare the various algorithms [17].

2.2. Image pre-processing

There is a lot of background noise in the input images that were collected from the dataset. The accuracy of the lung cancer identification system is hampered by the incorporation of various noises during the acquisition of images, including salt noise, a pepper noise, speckle noise and also Gaussian noise. Processing of images is employed to enhance visual information by correcting distortions that may have been introduced for the purposes of processing and analysis. Pre-processing is essential for improving image quality by adjusting exposure, sharpness, and de-noising. Many methods like the smoothing and the augmentation, are utilised to produce the desired result [18]. It smooths out the image by reducing noise and blemishes. In comparison to the input original image, filtered image appears to be crisper and clearer. The image has more contrast and dynamic range.

The median filter is a nonlinear method for cleaning up image data that has been contaminated by noise. It is predominantly used to eliminate the “salt as well as pepper” noise. When applied to an image, it replaces each pixel with the average value of those around it. The floor plan of this neighbour is called a window. The median value is found by arranging the window’s pixel values numerically and then use the median to replace them. Median filters are frequently used in digital image processing [19] due to their ability to preserve image edges, sharpness, and detail while simultaneously reducing noise. Hence, to increase the contrast of the grayscale image, a median filter was applied.

2.3. Image segmentation

Segmentation is the method of dividing images into comparable groups of image features. The thresholding technique is used in our proposed method. It is a method for segmenting input lung CT images in which information obtained from the grayscale image pixels that have a brightness that is either higher or lower than a threshold is discarded. It eliminates non-essential data from the image while preserving essential data [20]. The image histogram can be easily divided using a single threshold value, T. Segmenting an image requires pixel-by-pixel analysis, with each object and background pixel being labeled as foreground or background based on whether or not their grey level is above or below the T threshold. The effectiveness of this technique is determined by its histogram partitioning capability. The threshold was determined using an intuitive approach based on an analysis of said histogram using visual elements.

The algorithm for thresholding is shown below.

Step 1: Calculate the value of T.
Step 2: Segmentation by applying T: I G1, pixels brighter than T; I G2, pixels darker than (or equal to) T.
Step 3: Determine average intensities like m1 and m2 for G1 and G2.
Step 4: Compute new threshold value: Tnew=m1+m2.
Step 5: If |T−Tnew|>ΔT, back to step 2, otherwise stop.

2.4. Feature extraction

The feature extraction method is the widely used technique for extracting different features in the field of image processing. In order to understand an image, it is necessary to retrieve its edges and lines. During feature extraction from CT images of the lung, only pertinent information is collected to determine whether or not a region is cancerous. The characteristics of the binary image that endured are defined. Different features are collected from the segmented image using the GLCM technique. GLCM is a statistical method of the second order for texture analysis. The frequency of various combinations of grey levels co-occurring in an image or image section is calculated using the GLCM method. It is used to calculate texture features because its contents provide a metric for intensity variation at the target pixel. The image sections at any angle or offset are used to derive co-occurrence matrices. The GLCM matrix provides the basis for the computation of numerical features. Several significant features, including contrast, correlation, energy, homogeneity, entropy, kurtosis, and skewness, can be extracted from lung image segments. These characteristics are used to train machine learning algorithms to detect lung cancer [21].

2.5. Classification

Classification models are then trained on the extracted features to determine whether the input CT images are cancerous or normal. Different classification techniques, like SVM method, RF method, KNN, and the DT methods, are used in our proposed technique. A CSV file is used to store the features that were extracted using GLCM techniques. The feature database is divided into 70% training data and 30% testing data. The techniques are evaluated individually using quality metrics like an accuracy, a recall, F1-score, and a precision.

2.5.1. Support vector machine

SVM, which is a supervised method for machine learning, is used to resolve both the classification and also regression issues. Since a hyperplane is used to make the distinction, this binary classifier is deterministic rather than probabilistic. In addition to having a wide variety of applications, it is capable of solving linear as well as nonlinear problems. It begins by analyzing the data and then proceeds to classify the images. The SVM creates a hyperplane that partitions the available data into two or many classes. The SVM will begin by constructing a hyperplane that effectively partitions the classes before selecting the hyperplane that divides the classes in the most precise manner. In order to get the most out of the specific marginal plane, our primary goal should be to pick the hyperplane that has the best chance of having the greatest possible marginal distance [22]. The algorithmic design of the SVM that is used for CT image classification is depicted in Figure 2.

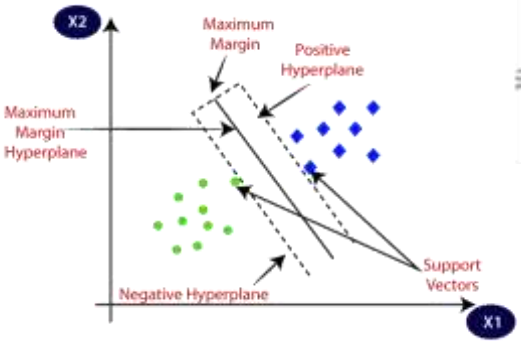


Figure 2. The structure of SVM classification algorithm for input CT lung images

2.5.2. Random forest

A random forest method classifies and predicts data in supervised machine learning. Random forests grow many decision trees, and the answer is based on their results. The dataset’s projected accuracy is increased by averaging many decision trees in its subgroups. Instead of using one decision tree, it gathers and analyses information from all of the trees, and it bases its output on the projections X that receive the majority of votes. The model becomes more precise as the forest grows, avoiding overfitting. It handles large databases with more variables. The ensemble model improves model accuracy and eliminates overfitting [23]. The following is an algorithmic representation of the random forest method.
Step 1: Consider n input sets from the provided dataset with k input sets.
Step 2: For each sample, a different decision tree is created.
Step 3: Each decision tree generates an output.
Step 4: The final result is determined by majority vote.

2.5.3. K-nearest neighbor

The KNN algorithm is a fundamental classification method that tries to find the most ideal match. This is an example of a technique for supervised classification. This technique differentiates between the database and the comparison dataset in order to achieve its desired outcomes. The score of the neighboring sample with the same value, which is the closest match, is used to determine the value that should be assigned to the test sample. When comparing the distances between research samples and database samples, various distance metrics, such as the Euclidean distance, the cosine distance, the similarity distance, and the city block distance, are used. It makes the assumption that the present data is similar to earlier cases and maintains the new incident in the group that is the most comparable. The KNN method retains data both while it is being trained and while it is classifying new data into similar groups. Although this method can be utilized for regressive analysis as well as classification, the latter application is where it shines the brightest. This approach is common because it is straightforward to comprehend and produces accurate results quickly. The KNN method takes N training vectors as input and determines the k points that are the closest neighbors to the point of interest, regardless of the labels [24]. The KNN algorithm for lung image classification is shown below:
Step 1: Find the K parameter, which is the number of closest neighbors.
Step 2: Determine the distance between each training example and the query instance.
Step 3: Based on the kth minimal distance, sort the distance and select the closest neighbors.
Step 4: Collect the closest neighbors in category Y.
Step 5: Use the category of the closest neighbor’s simple majority as the prediction value for the query instance.

2.5.4. Decision tree

The decision tree uses supervised learning to create a tree-shaped model from data (a group of nodes organized hierarchically). The entropy of the parent is measured and calculated first. The root node includes all the information, and the categorization goes on until the very end. The tree predicts results from new test data [25]. The classification using decision tree method for lung images is represented in Figure 3. The algorithm for decision tree method is explained:
Step 1: Get the list of attributes ready.
Step 2: Based on the attribute selection process, one of the attributes is chosen as a node.
Step 3: Perform the computations necessary to determine the information gain and the information gain ratio.

Step 4: Create the decision tree based on the number of values for the attribute.
Step 5: Separate the training dataset into multiple parts according to the total number of tuples, Di.
Step 6: Verify that the conditions for termination have been met.

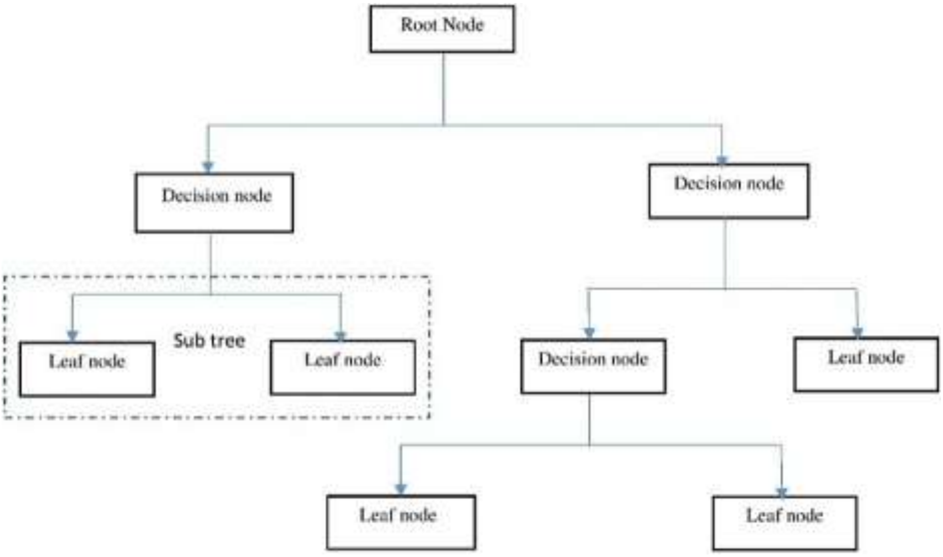


Figure 3. The organizational framework of the decision tree algorithm that was used for lung image classification

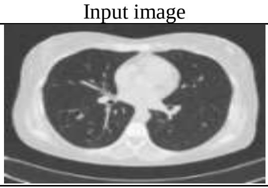
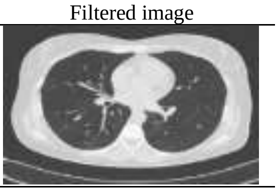
3. RESULTS AND DISCUSSION

Images of the lung taken from CT scans are taken from the LIDC database and used in the experiments. The database contains 4,000 images, each of which is 512 pixels wide and 512 pixels tall and is saved in the DICOM format. For the purposes of testing and training the model, the images are split into 70:30 ratios. In order to carry out the computation, the images that are in the DICOM format must first be converted into the png format.

3.1. Image preprocessing

A median filter removes CT scan noise and artifacts during image preprocessing. Gaussian noise, speckle noise, salt noise, pepper noise, and other types of noise are the main types of noise that can distort CT scan images. Using a median filter, noise is removed, thereby improving the image quality. The accuracy of image segmentation and feature extraction is improved by performing preprocessing on the image by removing undesired noise, which helps in accurate lung cancer detection. The image that was produced by applying the median filter to the input image is shown in Table 1.

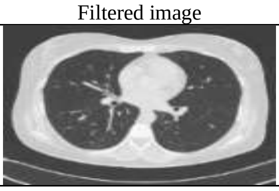
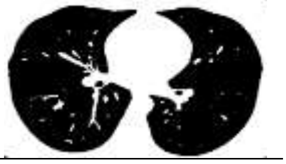
Table 1.Original image and filtered image

Image Id	Input image	Filtered image
Img-01		

3.2. Image segmentation

The thresholding method is one of the approaches that are utilized by the proposed approach. The process of thresholding, which is used to segment images of the lungs, includes eliminating information from pixels of a grayscale image if that pixel's value is greater than or less than a certain level. The image has been altered to eliminate any non-essential details while maintaining the critical information it originally had. The resultant image of the segmentation method is shown in the Table 2, which contains the output images.

Table 2.Segmented image from filtered image

Image Id	Filtered image	Segmented image
Img-02		

3.3. Feature extraction

The GLCM method is used to collect a variety of features from the image that has been segmented. Contrast, energy, correlation, the homogeneity, entropy, kurtosis, and the skewness are some of the essential characteristics that can be extracted from segmented lung images. These characteristics are fed into machine learning algorithms, which are then trained to make lung cancer predictions. According to Table 3, these features were extracted from 10 different lung input images by utilizing the GLCM method for the purpose of identifying lung tumors.

Table 3. The GLCM features generated from image segmentation

Image id	Contrast	Energy	Homogeneity	Correlation	Entropy	Kurtosis	Skewness
Img-01	0.0239	0.778	0.9905	0.9876	1.8581	8.2264	2.5541
Img-02	0.14	0.7955	0.9712	0.7969	1.9521	36.205	4.9288
Img-03	0.0867	0.9301	0.9855	0.6205	1.6402	76.116	6.9776
Img-04	0.0356	0.6423	0.9849	0.9881	2.7487	3.9335	1.6119
Img-05	0.1224	0.8133	0.9734	0.7636	2.0155	37.905	4.8762
Img-06	0.0137	0.9912	0.9979	0.4061	0.0621	951.73	26.915
Img-07	0.0827	0.4099	0.9733	0.9910	3.4922	1.7699	0.7502
Img-08	0.0167	0.9436	0.9941	0.7901	0.8200	41.951	5.3864
Img-09	0.1247	0.9085	0.9809	0.7785	2.0777	63.076	6.9327
Img-10	0.1334	0.8135	0.9726	0.7323	1.8885	37.363	4.8303

3.4. Classification

Classification models are trained using the features extracted from segmented image, to determine whether input lung CT images contain cancer or do not contain cancer. In the method that we have suggested, the GLCM features that have been extracted and saved in the CSV file are classified using a variety of classification methods like SVM, the KNN, RF, as well as DT. The different features in the database are split into two sections: the data used for training and testing data, with a ratio of 70% training data to 30% testing data. According to Table 4, each method is analyzed independently using various performance measures like an accuracy, a F1 score, a recall, as well as precision.

Table 4. The evaluation of various classification techniques using a variety of performance metrics

Methods	Accuracy	Precision	Recall	F1-Score
SVM	0.9932	0.99	0.99	0.99
Random Forest	0.9463	0.95	0.95	0.95
KNN	0.9248	0.92	0.92	0.92
Decision Tree	0.9328	0.94	0.93	0.93

The proposed methodology is conducted out on lung CT image datasets to accurately classify lung cancer. Texture features of lung CT images are extracted using a standard GLCM technique with various optimized novel features. This optimized resultant feature subset is then trained by various machine learning models to accurately predict and detect the type of lung cancer and compare it with other algorithms to determine the best performance model. In order to accurately detect and predict lung tumors, the researcher proposed a hybrid framework with various optimized machine-learning algorithms. In addition, the survival rate can be improved with the proposed method. Different performance metrics are utilised in order to conduct an effectiveness analysis of different algorithms on input CT images of the lungs in order to select the most appropriate algorithm for tumor detection. In the proposed method, four different non-parametric machine learning classifiers like SVM, RF, KNN and the DT methods have been hailed as the top classifiers for producing high accuracy, efficient experimental results, a training time reduction of over 50% for the proposed module, and obtained an accuracy of 99%.

Accuracy: it describes the number of correct predictions as well as predictions made overall, and is frequently used as the fundamental metric for evaluating models.

Accuracy=(TN+TP)/(TN+FN+TP+FP)

Precision: the number of accurate positive predictions made is one way to measure the precision of an analysis (true positives). The expression can be written as

Precision=TP/(FP+TP)

Recall: it is the percentage of positive cases correctly predicted by a classifier in comparison to total amount of positive samples in the dataset. This is also referred as sensitivity. The expression can be written as

Recall=TP/(TP+FN)

F1-Score: it is a measurement that takes into account both recall as well as precision. It is commonly known as the harmonic mean of the two.The traditional arithmetic mean is one method of calculating a "average" of values, but there is also a method known as the harmonic mean that can be used instead. The formula that is used to calculate F1-score is shown:

F1-Score=2*((Precision*Recall/(Precision+Recall))

The research is carried out using Python programming using appropriate inbuilt libraries and the tool used is ANACONDA.

Based on the experimental results, the SVM method achieved the highest possible accuracy, which was 99.32% in comparison to various other classification techniques. The accuracy rate of the decision tree method was measured at 93.288%, whereas random forest technique had an accuracy level of 94.63%. Based on the outcomes, the KNN technique achieved the lowest accuracy of 92.48%. Hence, it is concluded that the support vector machine accurately classifies whether the provided input CT images are normal or cancerous. Figure 4 illustrates the comparison of different classification methods in terms of accuracy. Based on the experimental results, the SVM method achieved the highest possible precision, which was 99% in comparison to various other classification techniques. The decision tree technique achieved a precision level of 94%, whereas the random forest method had a precision level of 95%. Based on the outcomes, the KNN technique achieved the lowest precision of 92%. Hence, it is concluded that the support vector machine accurately classifies whether the provided input CT images are normal or cancerous. Figure 5 illustrates the comparison of different kinds classification methods in terms of precision.

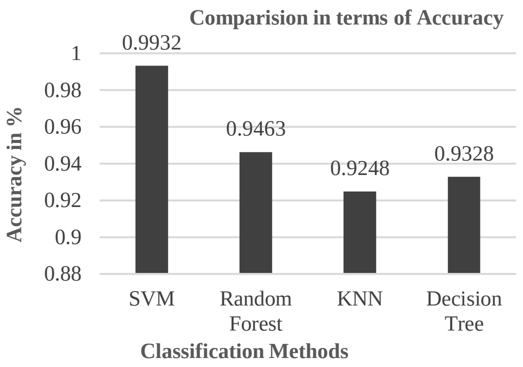


Figure 4. The comparison of various classification methods in terms of accuracy

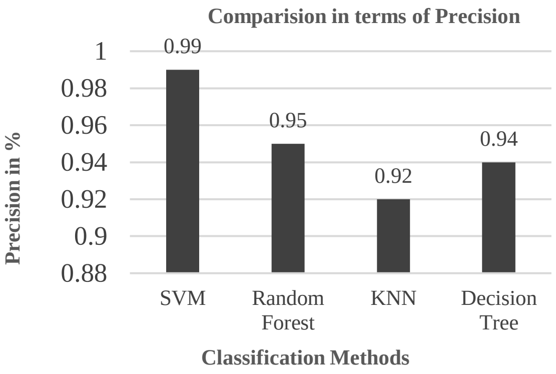


Figure 5. The comparison of various classification methods in terms of precision

Based on the experimental results, the SVM method achieved the highest possible recall, which was 99% in comparison to various other classification techniques. The recall rate for the decision tree technique was 93%, whereas random forest technique had a recall of 95%. Based on the outcomes, the KNN technique achieved the lowest recall of 92%. Hence, it is concluded that, the support vector machine accurately classifies whether the provided input CT images are normal or cancerous. Figure 6 illustrates the comparison of different classification methods in terms of recall. Based on the experimental results, the SVM method achieved the highest possible F1 score, which was 99% in comparison to various other classification techniques. The F1 score for the decision tree technique was 93%, whereas the random forest method had F1 score rate of 95%. Based on the outcomes, the KNN technique achieved the lowest F1 score of 92%. Hence, it is concluded that the support vector machine accurately classifies whether the provided input CT images are normal or cancerous. Figure 7 presents a comparison of several different classification techniques in terms of F1-score. Based on the results of the experiments, it is concluded that SVM identifies cancer in the lung with 99.32% accuracy in comparison to various other methods of classification.

The wiener filter and ANN technique for lung cancer detection had a detection accuracy of 94%, the super pixel density-based region approach and SVM classifier [16] for lung cancer detection had an accuracy of 88.23%, and the GLCM and SVM method had an accuracy of 83.33% [26]. All hybrid methods are tested on the same dataset in the proposed research work. Lung input images are collected from the standard LIDC database. A total of four thousand images in the 512*512 DICOM format with the necessary pre-processing are included in the database. The images are divided in a 70:30 ratio for training and training the model. The accuracy of the newly proposed method was 99.32%, which was higher than the accuracy of the existing method, which detected lung cancer more accurately. Table 5 represents the analysis results of this research work.

The research method cited in [26] uses GLCM and the SVM methods and achieved 83.33% of accuracy The proposed technique compares algorithms in each stage, uses standard GLCM for texture feature extraction, and uses best suited several hybrid machine learning techniques for accurate lung tumor classification by comparing each of the optimized classifiers using an accuracy, a recall, precision, and F1-score, thereby yielding an accuracy of 99.32% for detection of lung cancer more accurately.

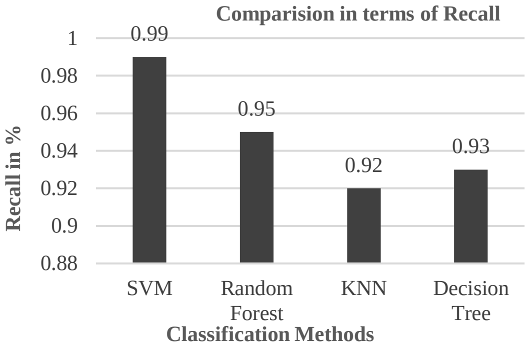


Figure 6. Comparing classification techniques based on recall

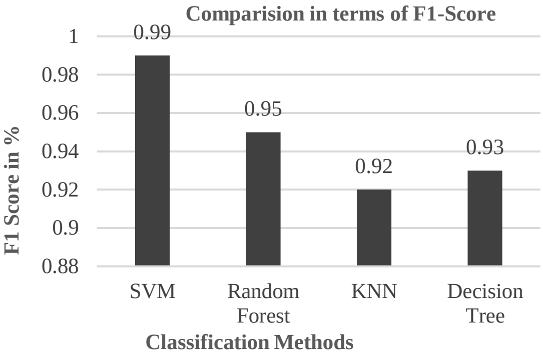


Figure 7. Comparing classification techniques based on F1-score

Table 5. Comparison of existing techniques with the proposed technique

Techniques	Accuracy
Halder *et.al* (2020)	88.23%
Firdaus *et.al* (2020)	83.33%
Manikandan *et.al* (2019)	94%
Proposed method	99.32%

4. CONCLUSION

Lung cancer is the second highest deadly disease, so an accurate diagnosis is crucial for improving treatment outcomes and survival for lung cancer patients. To find the best classification technique for the accurate detection of lung tumors, a number of machine learning techniques are implemented in the proposed work. The lung images used for the input are taken from the LIDC dataset.

The main contribution of this proposed work on research is to obtain a standard optimized model that helps in the accurate detection of cancer in lung. Today’s world is more concerned with accurate automated module identification systems. Therefore, the research work mainly focuses on improving image quality by using various pre-processing stages like image enhancement, segmentation, and feature extraction. CT scan images have been used for detection, and different automated module recognition systems have been proposed using techniques like image segmentation, the feature extraction, and pre-processing. The benefit of this approach is, it takes less storage space and increases the processing speed. This proposed work aims to detect and classify lung cancer precisely and accurately, based on various optimized machine-learning classifiers. The proposed research work gives better accuracy when compared to other existing techniques. Also, a comparative study is made on the CT image database to measure the performance metrics of proposed and existing techniques. The proposed module thus provides effective experimental results with reduced training time and classification accuracy of 99%. The Median filter to remove noise, the thresholding for segmentation, and GLCM for feature extraction are applied to the input images. Several different classification approachs, including the DT, the RF, the KNN as well as SVM, are applied to the features dataset and evaluated separately using performance metrics like an accuracy, a recall, the precision, and also F1-score. The experimental results clearly conclude that SVM technique has an accuracy rate of 99.32% for detecting lung cancer.

The results of further testing on a substantial amount of data will make it possible to incorporate new characteristics, such as the position, size, and texture of nodules. CNN is an effective method that can be used to improve accuracy. It is possible to determine the different stages of lung cancer based on factors such as age and gender.

ACKNOWLEDGEMENT

The researcher will like to express sincere thanks and appreciation to his esteemed mentor and supervisor for their ongoing inspiration, technical support, guidance, encouragement, and insightful suggestions that enabled him to undertake this research work.

REFERENCES

[1] B. T. Ahmed, “Lung cancer prediction and detection using image processing mechanisms: an overview,” *Signal and Image Processing Letters*, vol. 1, no. 3, 2019, doi: 10.31763/simple.v1i3.11.

[2] C. Shankara, S. A. Hariprasad, K. K. Pavan, Govardhan, and K. Basha, “A survey on lung cancer detection using convolution neural network in CT images,” *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 8, no. 7, pp. 449–452, 2021.

[3] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, “Prediction and classification of lung cancer using machine learning techniques,” *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012059.

[4] M. Jha, R. Gupta, and R. Saxena, “A review on non-invasive biosensors for early detection of lung cancer,” in *2020 6th International Conference on Signal Processing and Communication (ICSC)*, Mar. 2020, pp. 162–166, doi: 10.1109/ICSC48311.2020.9182775.

[5] D. M. Abdullah and N. S. Ahmed, “A review of most recent lung cancer detection techniques using machine learning,” *International Journal of Science and Business*, vol. 5, no. 3, 2021.

[6] S. S. Kanitkar, N. D. Thombare, and S. S. Lokhande, “Detection of lung cancer using marker-controlled watershed transform,” in *2015 International Conference on Pervasive Computing (ICPC)*, Jan. 2015, pp. 1–6, doi: 10.1109/PERVASIVE.2015.7087031.

[7] B. U. Dhaware and A. C. Pise, “Lung cancer detection using Bayasein classifier and FCM segmentation,” in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Sep. 2016, pp. 170–174, doi: 10.1109/ICACDOT.2016.7877572.

[8] R. Sathishkumar, K. Kalaiarasan, A. Prabhakaran, and M. Aravind, “Detection of lung cancer using SVM classifier and KNN algorithm,” in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Mar. 2019, pp. 1–7, doi: 10.1109/ICSCAN.2019.8878774.

[9] L. Patil, A. Sirsat, D. Kamble, and M. Y. Pawar, “Lung cancer detection using decision tree algorithm,” *International Research Journal of Engineering and Technology*, vol. 4, no. 2, pp. 1885–1888, 2017.

[10] D. Jayaraj and S. Sathiamoorthy, “Random forest based classification model for lung cancer prediction on computer tomography images,” in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Nov. 2019, pp. 100–104, doi: 10.1109/ICSSIT46314.2019.8987772.

[11] B. Abdillah, A. Bustamam, and D. Sarwinda, “Image processing based detection of lung cancer on CT scan images,” *Journal of Physics: Conference Series*, vol. 893, Oct. 2017, doi: 10.1088/1742-6596/893/1/012063.

[12] S. Kalaivani, P. Chatterjee, S. Juyal, and R. Gupta, “Lung cancer detection using digital image processing and artificial neural networks,” in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Apr. 2017, pp. 100–103, doi: 10.1109/ICECA.2017.8212773.

[13] P. Lobo and S. Guruprasad, “Classification and segmentation techniques for detection of lung cancer from CT images,” in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jul. 2018, pp. 1014–1019, doi: 10.1109/ICIRCA.2018.8597273.

[14] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, “Lung cancer detection using CT scan images,” *Procedia Computer Science*, vol. 125, pp. 107–114, 2018, doi: 10.1016/j.procs.2017.12.016.

[15] T. Manikandan, B. Devi, and T. Helanvidhya, “A computer-aided diagnosis system for lung cancer detection with automatic region growing, multistage feature selection and neural network classifier,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 1S, pp. 409–413, Dec. 2019, doi: 10.35940/ijitee.A1081.1191S19.

[16] A. Halder, S. Chatterjee, and D. Dey, “Superpixel and density based region segmentation algorithm for lung nodule detection,” in *2020 IEEE Calcutta Conference (CALCON)*, Feb. 2020, pp. 511–515, doi: 10.1109/CALCON49167.2020.9106569.

[17] H. Wang, C. Wu, J. Chi, X. Yu, and Q. Hu, “Speckle noise removal in ultrasound images with stationary wavelet transform and canny operator,” in *2019 Chinese Control Conference (CCC)*, Jul. 2019, pp. 7822–7827, doi: 10.23919/ChiCC.2019.8866685.

[18] C. Shankara, S. A. Hariprasad, and D. U. Latha, “Detection of lung cancer using convolution neural network,” *SN Computer Science*, vol. 4, no. 3, 2023, doi: 10.1007/s42979-022-01630-y.

[19] A. Shah *et al.*, “Comparative analysis of median filter and its variants for removal of impulse noise from gray scale images,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 505–519, Mar. 2022, doi: 10.1016/j.jksuci.2020.03.007.

[20] C. Shankara and S. A. Hariprasad, “Artificial neural network for lung cancer detection using CT images,” *International journal of health sciences*, pp. 2708–2724, Apr. 2022, doi: 10.53730/ijhs.v6nS2.5639.

[21] P. K. Mall, P. K. Singh, and D. Yadav, “GLCM based feature extraction and medical X-RAY image classification using machine learning techniques,” in *2019 IEEE Conference on Information and Communication Technology*, Dec. 2019, pp. 1–6, doi: 10.1109/CICT48419.2019.9066263.

[22] B. Gaye, D. Zhang, and A. Wulamu, “Improvement of support vector machine algorithm in big data background,” *Mathematical Problems in Engineering*, pp. 1–9, Jun. 2021, doi: 10.1155/2021/5594899.

[23] Z. Bingzhen, Q. Xiaoming, Y. Hemeng, and Z. Zhubo, “A random forest classification model for transmission line image processing,” in *2020 15th International Conference on Computer Science and Education (ICCSE)*, Aug. 2020, pp. 613–617, doi: 10.1109/ICCSE49874.2020.9201900.

[24] P. Nair and I. Kashyap, “Classification of medical image data using k nearest neighbor and finding the optimal k value,” *International Journal of Scientific and Technology Research*, vol. 9, no. 4, pp. 221–226, 2020.

[25] Z. I. Alassar, “Decision tree as an image classification technique.” Department of City and Regional Planning, Faculty of Architecture, Akdeniz University, pp. 1–7, 2020.

[26] Q. Firdaus, R. Sigit, T. Harsono, and A. Anwar, “Lung cancer detection based on CT-scan images with detection features using gray level co-occurrence matrix (GLCM) and support vector machine (SVM) methods,” in *2020 International Electronics Symposium (IES)*, Sep. 2020, pp. 643–648, doi: 10.1109/IES50839.2020.9231663.

BIOGRAPHIES OF AUTHORS

Shankara Chikkalingaiah is working as a lecturer in the Department of Electronics and Communication Engineering at Government Polytechnic Nagamangala Mandya and currently, he is deputed for a full-time Ph.D. at Jain University Bangalore. He completed his M.Tech degree in Digital Communication and Networking at SJBIT Bangalore affiliated with Visveswaraya Technological Univerity Belgavi. He has more than 10 years of teaching experience and his research areas are Medical image processing and Machine Learning. He has published 4 research publications in International Journals and presented 3 papers at National and International conferences. He can be contacted at email: shankara151987@gmail.com.

Subbarao Anantha Padmanabha Rao Hari Prasad is currently working as Director at the Faculty of engineering and technology, Jain University Bangalore. He completed his B.E degree in the year 1991, and his M.E in the year 2000, He Completed his Ph.D. degree in the year 2011 and two additional highest degrees, doctor of science degrees in the year 2013 and 2014 for post-doctoral research work in communication and embedded systems. Working with the Department of Electronics and Communication Engineering, Jain University, Bangalore. He also established industry-supported labs and served as visiting professor for various reputed colleges in Bangalore. He executed funded projects tuning to 1.5 crore rupees for various government funding agencies. Guided 35 PG projects, 40 UG projects, and 7 students who completed Ph.D. and currently guiding 8 Ph.D. students. He has delivered expert talks at conferences and workshops in the area of embedded systems and Microwave Engineering domains. Worked as main Coordinator as well as a mentor for input/output and outcome Based Education Accreditation models. Won awards for best teacher for publishing international conference papers and textbooks. He has published 69 research publications in both International and National Journals and presented 35 papers at National and International conferences. He has written a textbook on Advance Microprocessor and reviewed 3 books. He can be contacted at email: sa.hariprasad@jainunversity.ac.in.

Latha Dabbegatta Uggregowda is currently working as Assistant Professor in Department of CSE (AI and ML) at Vidyavardhaka College of Engineering, Mysuru. She is pursuing a part time Ph.D. at Jain University Bangalore. She completed her B. E from University Visveswaraya College of Engineering in Information Science and M.Tech from MIT, Mysuru affiliated with Visveswaraya Technological University, Belgavi. She has 2 years of Industrial experience as Project Engineer at Wipro Technologies, Bangalore and 6 years of teaching experience and her research interests are Image processing, machine learning and Deep learning. She can be contacted at email: lathadu@vvce.ac.in.