# Lung Nodule Classification using Transfer Learning and Morphological Feature Extraction

Nafe Muhtasim Hye[1], Umma Hany[2], Tahmina Islam[3], Nusrat Nawreen[4], Abdullah Al Mamun[5]

September 2, 2022

[1,2,3,4,5]Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.
email:[1]nafemuhtasim98@gmail.com,[2]uhany.eee@aust.edu,[3]tahminaislam837@gmail.com,
[4]nahinnawreen@gmail.com,[5]abdullah78675645@gmail.com

**Abstract**

Lung Cancer is an uncontrolled growth of tissue that causes a lump in human lung that we call lung tumor or nodules. Lung cancer does not show any symptoms in early stage. If the lung cancer can be detected at early stage, it can increase the survival rate. In this paper, we propose a novel approach of lung nodules detection and classification by processing the Computed Tomography (CT) scan images. We apply different pre-processing steps for resizing, smoothing and enhancement of the data images. Then, we propose the morphological segmentation and feature extraction methods for segmentation of the target regions of the nodules and to extract the features. We also apply transfer learning based feature extraction using VGG16 model. Finally, the preprocessed data and the extracted features are trained and classified to three labels using different deep learning and machine learning algorithms. The methods are simulated and the accuracy is compared. It is found that the proposed models shows significantly high accuracy in detection and classification of the lung nodules with high computational intelligence.

Keywords: Lung cancer detection, Transfer learning, VGG16, Morphological feature extraction.

## 1 Introduction

Lung cancer is generally a malignant tumor. People accompanying pleura disease to a large degree like emphysema and premature chest problems have more chance to be diagnosed with body part malignancy. Overuse of tobacco, cigarettes and secondhand smoking are the major risk factor that leads to bronchi malignancy. The size of the tumor and the extent of it's spreading determines the stage of the tumor [1]. Lung Tumor can be classified into two types; Benign and Malignant. A tumor that is no threat to invading other tissues is considered Benign. When the cell is abnormal and can grow uncontrollably and spread to other parts of the body, is cancerous or Malignant tumor. This spreading process is called metastasis [2]. Lung disease can be seen on chest radiographs and computed tomography (CT) scan images. Among these, CT scan images is the most reliable and effective method as it shows a detailed picture of the object and it's growth [3]. In the literature, there are several works on the detection and classification of lung diseases.

In paper [4], the author proposes various machine learning methods for detecting lung cancer nodule from CT scan images and apply the noise removal method for better accuracy. The reported results in [4] indicate that Artificial Neural Network (ANN) gives 82.43% accuracy with image processing and the Decision tree gives 93.24% accuracy without image processing. In paper [5], the authors propose lung cancer detection in Positron Emission Tomography/Computed

Tomography (PET/CT) images by using texture and fractal feature descriptors classified by machine learning and artificial intelligence techniques. As per their report, it outperforms with 98.10% accuracy using Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. In [6], a DenseNet non-negative sparse and collaborative representation (DenseNet-NSCR) method is proposed for benign and malignant classification of the lung tumors. In this paper, the data are trained in a DenseNet to extract the feature vector of the full connection layer. They reported better robustness and generalization capabilities with 99.10% accuracy. In paper [7], the authors evaluate several available machine learning algorithms in the literature which can be used for lung cancer detection associated with IoT devices. In paper [8], the authors propose a parallel deep learning model with a hybrid attention mechanism for image segmentation and apply the DenseNet module. The results demonstrate that the parallel deep learning with hybrid attention mechanism perform well in image segmentation of lung cancers and the accuracy can reach 94.61%. In paper [9], segmented tumor images are fed into a hybrid feature detection and extraction model called Maximally Stable Extremal Regions and Speeded Up Robust Features (MSER-SURF). Then, the extracted MSER-SURF features are classified using 1D CNN model. The reported accuracy in [9] is 96 ± 3%. In paper [10], machine learning algorithms such as the Multi-layer Perceptron (MLP), K-Nearest Neighbor (KNN) and SVM classifier are used to classify the extracted features using Gray Level Co-occurrence Matrix (GLCM). The reported accuracy using MLP is 98% where SVM acquires 70.45% accuracy and the acquired accuracy using KNN is 99.2%. In paper [11], the authors compare the accuracy of lung cancer classification of SVM, KNN and Convolutional Neural Network(CNN) by WEKA tool at the beginning phase of malignancy. The resulted outcome shows that SVM gives the best outcome with 95.56%, CNN with 92.11% and KNN with 88.40%. In paper [12], the authors report several works on image segmentation, feature extraction as well as various techniques to classify and detect lung cancer at early stage. In paper [13], the authors propose an ingenious methodology of lung tumor detection using image segmentation for feature extraction and machine learning-based classification to classify the test images of lung nodules into normal and abnormal. In paper [14], the author classifies lung cancer images into two classes (benign and malignant) using fuzzy logic accompanied by region of interest (ROI) extraction. SURF technique and Genetic algorithm (GA) are used in [14] for feature extraction and optimization and then SVM with Feed-Forward Back Propagation Neural Network (FFBPNN) is used to create a hybrid classification algorithm reducing the computational complexity. The authors report overall classification accuracy of 98.08% in [14].

In this paper, we propose novel approaches of lung cancer detection and prediction by applying machine learning and deep learning classification on the CT scan images. First, we apply pre-processing steps to resize and enhance the images. Then, the preprocessed raw data are applied to the classifiers. However, as the raw data requires high processing time, we apply VGG16 model for transfer learning (TL) based feature extraction. The extracted TL features are then applied to the classification to reduce the processing time. For real time health care and improved computational intelligence, we propose the methods of image segmentation for different types of nodules using thresholding and morphological operations. Next, we propose the morphological feature extraction to extract distinct features of the nodules. Finally, the low dimensional morphological features are applied for classification which significantly improve the accuracy with low computational complexity and high computational intelligence with with low processing time. We apply different deep learning and machine learning based classifiers to train the images and to classify the nodules into three labels; normal, benign and malignant. Among the classifiers, the deep learning based convolutional neural network (CNN) shows 98.19% accuracy with preprocessed raw data. Among the machine learning classifiers, logistic Regression (LR) shows 100% accuracy using both preprocessed raw data and transfer learning features. With the extracted morphological features, most of machine learning algorithms show significantly high accuracy and SVM show the highest accuracy of 99.71%. We compare the training time in seconds to analyze the computational intelligence of the proposed systems. As we can see

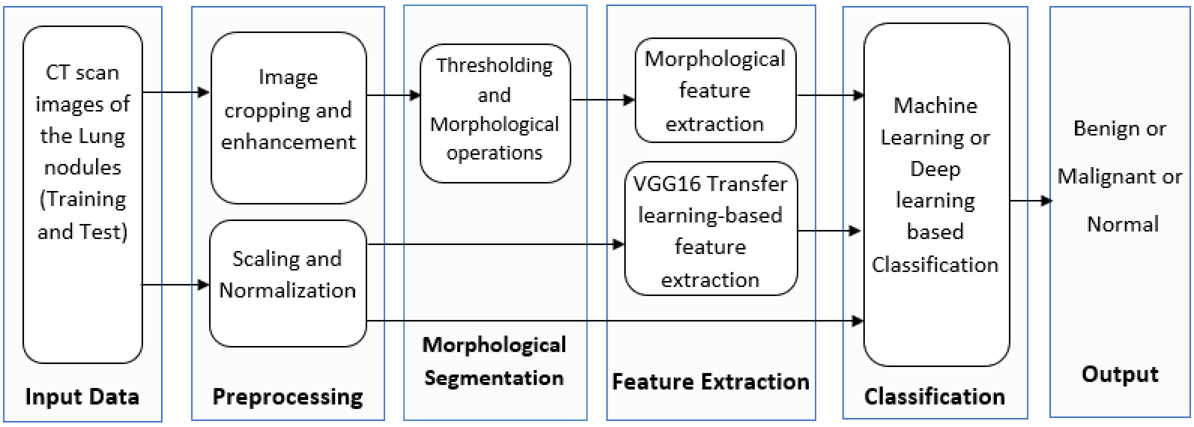

CT scan images of the Lung nodules (Training and Test)
Input Data
Image cropping and enhancement
Scaling and Normalization
Preprocessing
Thresholding and Morphological operations
Morphological Segmentation
Morphological feature extraction
VGG16 Transfer learning-based feature extraction
Feature Extraction
Machine Learning or Deep learning based Classification
Classification
Benign or Malignant or Normal
Output

Figure 1: System Architecture

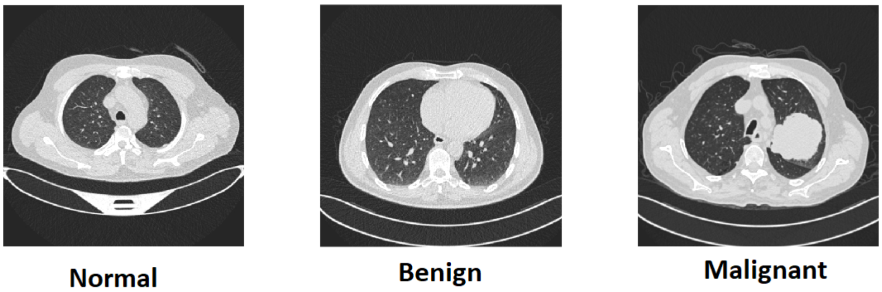
that morphological feature extraction reduces the dimension of the data significantly resulting in very low computational complexity and processing time which increases the computational intelligence of the proposed systems. Thus, it can be concluded the proposed morphological feature extraction shows high accuracy with low computational complexity and high intelligence. We also compare the accuracy with other methods in the literature and find that the proposed methods outperforms the literature with high accuracy and high computational intelligence.

## 2 System Architecture

The lung nodules are detected and classified using the system as shown in Figure 1. The system starts with lung image acquisition using CT scanner. The collected CT scan images are loaded to the system for preprocessing. We propose two methods of data preprocessing by using Scaling and normalization and by using Data cropping and enhancement. The processed images are forwarded to the feature extraction modules. The extracted features of the processed data are then trained and tested by the classification modules to predict the class of the lung nodules. The lung nodules are classified both by using machine learning and deep learning methods. In our proposed approach as shown in Figure 1, we classify the lung nodules using three approaches to compare the accuracy of classification. In the first approach, the classifier receives the raw data preprocessed by the proposed scaling and normalization method. In the second approach, the features of the preprocessed data are extracted using VGG16 based transfer learning and classified using machine learning or deep learning methods. In the third approach, we propose data preprocessing, morphological segmentation and morphological feature extraction methods before applying to the classifier to reduce computational complexity and to improve the processing time. Then, the extracted features are trained and classified with the available machine learning models. Finally, the CT scan images of the lung nodules are classified into three classes Benign, Malignant or Normal. The predicted label or class of the lung nodules are then compared to the original label to verify the accuracy. The detail methodologies are explained in the following section.

## 3 Image Segmentation and Feature Extraction

The features of the preprocessed CT scan images of the lung nodules are extracted using two methods; (i) Transfer learning (TL) based feature extraction and, (ii) Proposed Morphological segmentation and feature (MSF) extraction. The images are preprocessed before feature extraction. For morphological feature extraction, morphological image segmentation methods are applied to separate the region of interest (ROI) of the lung nodules. Then, the preprocessed data



Normal Benign Malignant

Figure 2: CT scan images of lung nodules of three classes

or the extracted features are trained and classified using different deep learning and machine learning algorithm.

## 3.1 Data preprocessing

First, the dataset of CT images of lung nodules are splitted into training and test images. 80% of the data are used for training and 20% are used for testing. The dataset contains CT scan images of normal, benign and malignant type nodules. Figure 2 shows the samples of three classes of lung nodules. The training and test data are preprocessed using the following two proposed methods.

### 3.1.1 Method 1: Scaling and normalization

The CT scan data is preprocessed using scaling and normalization before applying to the classifier. The data is also preprocessed by Method 1 for transfer learning based feature extraction. In this method, first the CT scan data in jpg format are loaded in BGR color space. Then, the BGR color format is interpreted and converted to RGB format. We apply scaling to transform data so that it fits within a specific scale and to change the range of data. The normalization is applied to change the observations so that they can be described as a normal distribution. We randomly shuffled the train pictures into a state of 25. Then we scale each pixel using a factor of 255. The majority of picture data has integer pixel values between 0 and 255. Small weight values are processed by neural networks, while high integer values might interfere with or slow down learning. Since, each pixel value of the image should range from 0 to 1, normalizing the pixel values is a good option. We use min-max scaling as follows to normalize the data such that the features value remains within a specific range 0 and 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

where, $x'$ is the normalized value. We divide all pixel values by 255, which is the biggest pixel value. Regardless of the actual range of pixel values that are present in the image, this is done across all channels.

### 3.1.2 Method 2: Image cropping and enhancement

This method is applied for morphological feature extraction. It includes data loading, image resizing, cropping and enhancement processes. First, all the the collected CT scan images are loaded and converted to grayscale images. Then, the images are resized and cropped to get the required dimensions. Finally, the contrast level of the images are adjusted to enhance the image features such as boundaries and edges. The original image, the cropped image and the enhanced image are shown in Fig. 3. After applying the image cropping and enhancement process, the raw RGB image of 512×512×3 uint8 dimension is converted to binary image of 136×151 logical dimension.

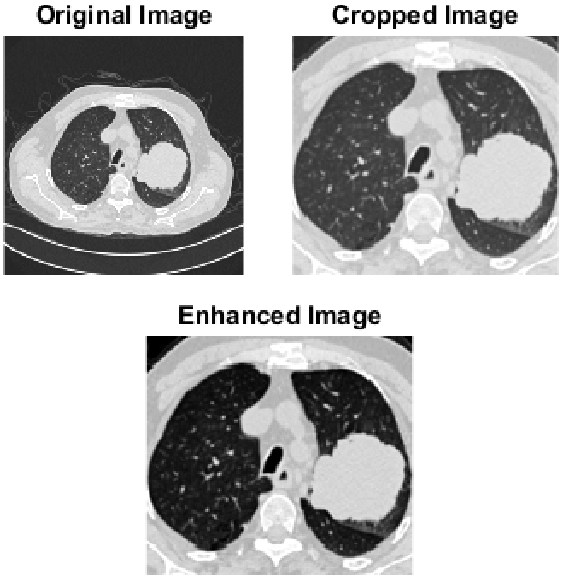


Figure 3: Image enhancement

## 3.2 Morphological segmentation

For morphological segmentation, the CT scan images are preprocessed first using image enhancement processes as discussed in the previous subsection. Then, the segmentation processes are applied for partitioning an image into multiple segments to locate the region of interest (ROI) of the lung. The ROI is required to be segmented to extract distinct features of different classes of lung nodules. The ROI is segmented using thresholding and morphological operations as follows.

### 3.2.1 Thresholding

Thresholding is the simplest method of image segmentation and a way to convert the gray-scale image to binary image based on the computed threshold value. The threshold level indicates the intensity value of the image. There are several thresholding methods including the maximum entropy method, balanced histogram thresholding, Otsu's method (maximum variance), and k-means clustering. We apply Otsu's method [15] to compute the global threshold of the contrasted grayscale image. In the binary conversion, the output binary image replaces all pixels in the input grayscale with luminance greater than level with the value 1 (white) and replaces all other pixels with the value 0 (black).

### 3.2.2 Morphological operations

We apply the following morphological operations on the thresholded binary images for the segmentation of the ROI of malignant, benign and normal nodules and to remove the unnecessary features.

(i) Creating structuring element: First, we extract the background of the binary images by creating a morphological structuring element (SE) of disk shape.

(ii) Opening operation: Then we apply morphological opening operation to open the SE from the binary image and to get the background of the image.

(iii) Background subtraction: The extracted background is then subtracted from the binary image to get the background subtracted image.

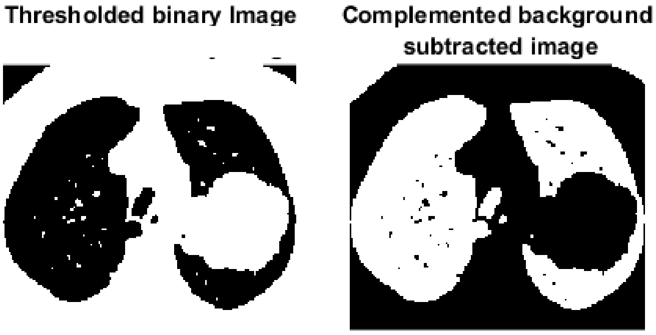


Figure 4: ROI segmentation of the Malignant type nodule

(iv) Complement image: The background subtracted images are complemented. The complemented background subtracted binary image is considered as the ROI of the malignant type nodules. Therefore, no further morphological operations are applied on the images if malignancy is detected.

(v) Image fill operation: The image fill operation is applied to fill the holes in the complemented background subtracted binary images.

(vi) Logical XOR operation and complement: The logical exclusive OR operation is applied on the complemented images (step iv) and the holes filled images (step v) to obtain the ROI of the benign type nodules. The XOR operated image is complemented to get the ROI of the normal type nodules.

Figures 4, 5 and 6 show the segmentation processes to get the ROI of the malignant, benign and normal type nodules. Fig. 4 shows the thresholded binary image and the complemented background subtracted ROI of a malignant type nodule. Fig. 5 shows the thresholded binary image, the complemented background subtracted image, the holes filled image and the XOR operated ROI of a benign type nodule. Fig. 6 shows the thresholded binary image, the complemented background subtracted image, the holes filled image and the complemented XOR operated ROI of a normal type nodule.

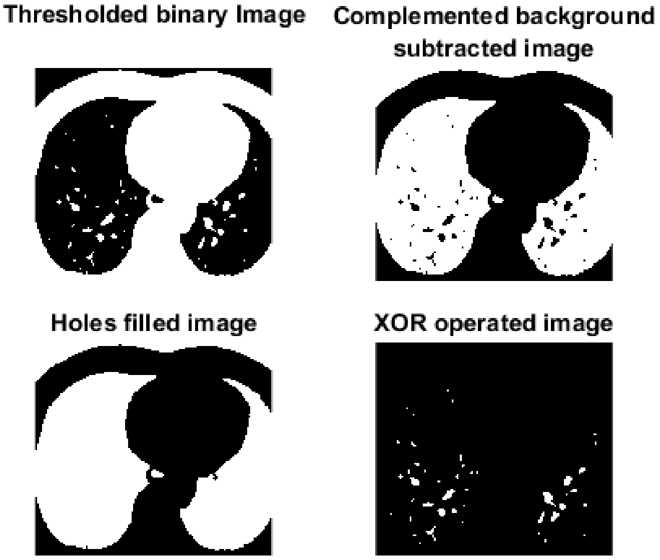


Figure 5: ROI Segmentation of the Benign type nodule

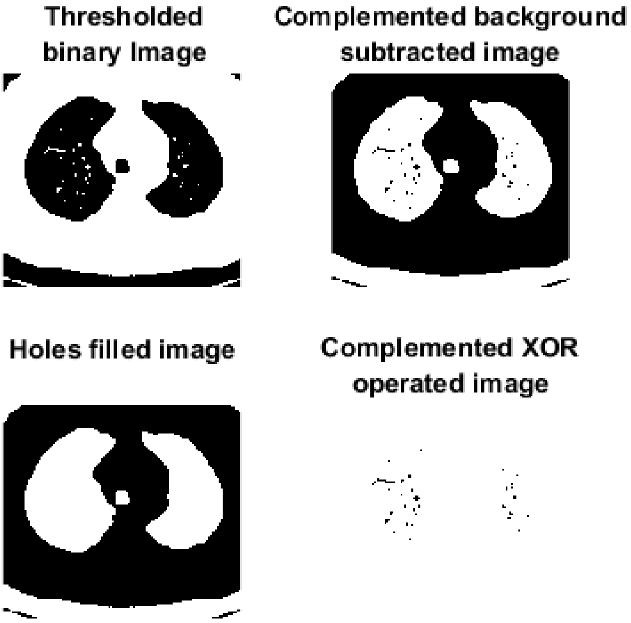

Thresholded binary Image
Complemented background subtracted image
Holes filled image
Complemented XOR operated image

Figure 6: ROI Segmentation of the Normal type nodule

## 3.3 Feature extraction

The features of the segmented ROI of the lung nodules are extracted using both morphological feature extraction and transfer learning based VGG16 feature extraction. The methods are explained as below.

### 3.3.1 Morphological feature extraction

In the morphological feature extraction module, the features of the segmented ROI of malignant, benign or normal type nodules are extracted by measuring the following four regional properties:

(i) Area: It is a measure of the actual number of pixels in the region.

(ii) Eccentricity: Considering an ellipse having equivalent second-moments as the region, Eccentricity is measured as the ratio of the distance between the foci of the ellipse and the length of it's major axis. This value is also referred to as the irregularity index of circularity or roundness. The value is between 0 and 1. An ellipse with 0 eccentricity can be defined as a circle and the ellipse can be defined as line segment if the eccentricity is 1.

(iii) Perimeter: This is the distance around the boundary of the region which is measured by the number of pixels in the boundary of the region. The perimeter is calculated as

$$\text{Perimeter} = \sum_{i=}^{N-1} d_i \tag{2}$$

(iv) Compactness: Compactness is defined as the ratio of the area of an object to the area of a circle with the same perimeter. Benign tumors are more smooth and round shape than a malignant tumor. Therefore, we use the following alternate formula of compactness

$$\text{Compactness} = \frac{\text{Perimeter}^2}{2 \times \text{Area}} \tag{3}$$

By applying morphological feature extraction, the dimension of the extracted features is significantly reduced as compared to the dimension of the original raw data. The dimension of the extracted features is 4×1 double containing the four extracted features of each segmented regions. Thus, it compresses the size of the training and test data significantly which results in less computational complexity and reduced training and test time of classification.

### 3.3.2 Transfer learning for feature extraction

When a trained model is repurposed for a different related task using machine learning approach is known as transfer learning. Transfer learning is an optimization that enables quick advancement or increased performance. There are many transfer learning methods such as VGG16, VGG19 etc. We use the VGG16 model as shown in Fig. 7 for feature extraction. VGG16 model has 16 convolutional layers including the Maxpooling layers, 3 dense layers (2 fully connected layers and 1 SoftMax classifier) and an output layer of 1,000 nodes [16]. For feature extraction, a new dataset is created from the input image dataset of lung nodules using the pre-trained model. The detail steps are as follows:

(i) First, the convolutional and pooling layers are imported and the fully-connected dense layers at the "top" of the model are removed.

(ii) Then, the image data are passed through the pre-trained layers to extract the visual features. Since VGG16 model is trained on millions of photos, the convolutional layers can recognize the generic visual features of the images with the trained weights and provides a feature stack containing the recognized visual features.

(iii) The 3 dimensional feature stack is then flattened before applying to other classifiers.

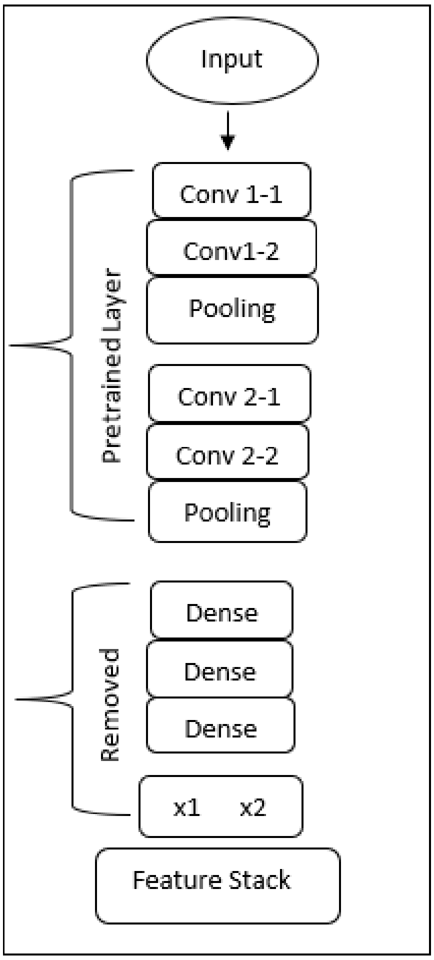


Figure 7: VGG16 model for feature extraction

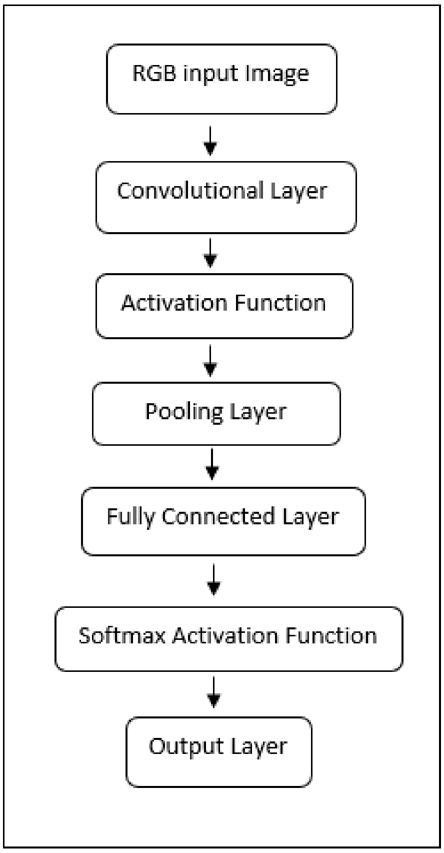


Figure 8: Convolutional Neural Network Architecture

# 4 Classification

The preprocessed raw data or the extracted features of the lung nodules are applied to the deep learning or machine learning based classifiers for classification. The different types of classifiers are briefly explained below.

## 4.1 Deep learning classifier

The Convolutional Neural Network (CNN) architecture of the deep learning model as shown in Fig. 8 is used for classification. Prior to deep learning, we add two convolutional layers with an input layer of (224, 224, 3) and an element-wise activation function on each of those, commonly known as a Rectified-Linear Unit (ReLu). The ability of activation of the input nodes is decided by the ReLu layer. If the filters in the convolution layer picks up a visual characteristic, the activation is indicated. ReLu function operates by applying a max (0, x) function thresholded at 0. Followed by two Maxpooling layers of (2, 2), a down-sampling strategy is applied to reduce the width and height of the output volume. After adding a flattened layer, two dense layers with 128 neurons and 3 neurons output, respectively are added. The model is then trained and tested using Google Colaboratory GPU. We chose the optimizer Adam and the sparse categorical cross entropy loss function with batch size = 64, epoch = 20 for the compilation stage to optimize the model during training and to minimize the loss function.

## 4.2 Machine learning classifiers

In our proposed system, we apply the following supervised Machine Learning algorithms on preprocessed raw data, transfer learning features and on the morphological features. In supervised machine learning, the model is trained using algorithms to discover patterns in a dataset of features and labels, and the model is then used to predict the labels on the features of a new dataset.

**1. Decision Tree (DT):** In Decision Tree classification algorithm, internal nodes represent dataset attributes, branches represent decision rules and each leaf node provides the output of the decision in the tree-structured classifier [17].

**2. k-Nearest Neighbors (KNN):** The K-Nearest Neighbour method assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories. The K-NN method stores all available data and classifies a new data point based on its similarity to the existing data [18].

**3. Random Forest (RF):** Random forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random forest correct the decision trees' habit of overfitting to their training set [19].

**4. Extra Trees (ET):** The Extra Trees classifier, also known as Extremely Randomized Trees, is comparable to a Random Forest classifier [20]. We alter our approach to tree construction in order to add more variance to the ensemble. The following standards are used to build each decision stump:

i) Each stump is constructed using the training set's data.

ii) The best split is found by searching through a subset of randomly chosen features to construct the root node or any node (number of features). Each feature's split is determined randomly.

iii) The decision stump can go only as deep as one.

**5. Extreme Gradient Boosting (XGB):** Extreme Gradient Boosting (XGB) is a supervised, distributed, scalable gradient-boosted decision tree (GBDT) machine learning framework. For classification and regression, a GBDT is a decision tree ensemble learning approach similar to a random forest [21].

**6. Support Vector Machine (SVM):** Support Vector Machine classifier finds the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category. A hyperplane is the optimal choice boundary which is created by the support vectors [22].

**7. Logistic Regression (LR):** Logistic regression predicts a categorical dependent variable from a set of independent variables. It can generate probabilities and classify new data using both continuous and discrete datasets [23].

# 5 Performance Evaluation

We apply the following performance metrics to evaluate the performance of the classification models [24].

## 5.1 Confusion matrix

Confusion Matrix is a visualization of ground-truth labels versus model predictions. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class. Each cell in the confusion matrix represents any of the following evaluation factors:

1. True Positive (TP) signifies how many positive class samples are predicted correctly.
2. True Negative (TN) signifies how many negative class samples are predicted correctly.
3. False Positive (FP) signifies how many negative class samples are predicted incorrectly.
4. False Negative (FN) signifies how many positive class samples are predicted incorrectly.

### 5.2 Precision

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of true and false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

### 5.3 Recall

Recall is the ability of a classifier to find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

### 5.4 F1 score

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

$$\text{F1 score} = \frac{2\ \times \text{Recall}\ \times \text{Precision}}{(\text{Recall} + \text{Precision})} \tag{6}$$

### 5.5 Accuracy

Classification accuracy is defined as the number of correct predictions divided by the total number of predictions, multiplied by 100. The classification or test accuracy is calculated as

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \times 100 \tag{7}$$

### 5.6 Cross entropy loss function

In classification model, the probability of each predicted class is compared to the desired output 0 or 1 of actual class. A loss function is calculated based on the probability of how far it is from the actual expected value. The penalty is logarithmic resulted in large score for large differences close to 1 and small score for small differences tending to 0. Cross-entropy loss is used to optimize the model by adjusting the weights during training to minimize the loss tends to 0. Cross-entropy loss function is defined as [25]

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i), \quad \text{for n classes} \tag{8}$$

where, n is the number of classes, $t_i$ is the true label and $p_i$ is the Softmax probability for the $i^{th}$ class.

## 6 Simulation and Results

We collect dataset of 1097 CT scan images of three types of lung nodules in jpg format from the Kaggle database [26]. The database contains an assortment of Computed Tomography (CT) images of patients with and without lung diseases. The dataset contains 120 CT scan images of benign class, 561 of malignant class, and 416 of normal class. The proportion of each classes of data is given in Fig. 9.

First, we simulate the preprocessed raw dataset using deep learning and machine learning based classifiers for lung nodule detection and classification. The image data are preprocessed using our proposed methods for resizing, smoothing and enhancement. The resulted preprocessed data are shown in Fig. 3. For classification, 80% of the dataset are trained with known label of different type of nodules and 20% of the dataset are tested. The data are splitted as shown in Table 1. Then, the Convolutional Neural Network (CNN) architecture of the deep learning model is used for classification of lung nodules using preprocessed raw dataset. The training and test accuracy of the CNN classifier is analyzed for increasing number of epochs. Figures 10 and 11 shows the accuracy curve and the loss curve of CNN for increasing number of epochs using raw dataset. It is observed in Fig. 10 and Fig. 11 that both training and test accuracy improves with the increasing number of epochs and both the training and test loss decreases with the increasing epochs. Table 2 summarizes the accuracy scores of the deep learning CNN based classification approach and it is observed the CNN shows 98.19% test accuracy using preprocessed raw data.

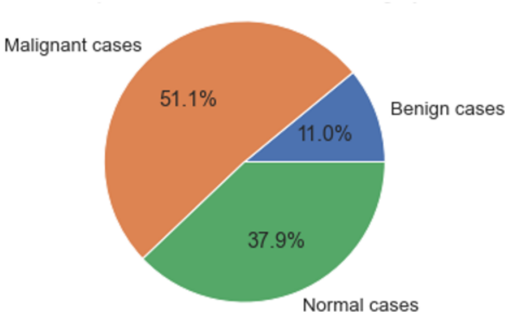


Figure 9: Proportion of different classes of lung nodule images

Table 1: Data Splitting

| Classes | Train | Test |
|---|---|---|
| Benign | 96 | 24 |
| Malignant | 448 | 113 |
| Normal | 332 | 84 |

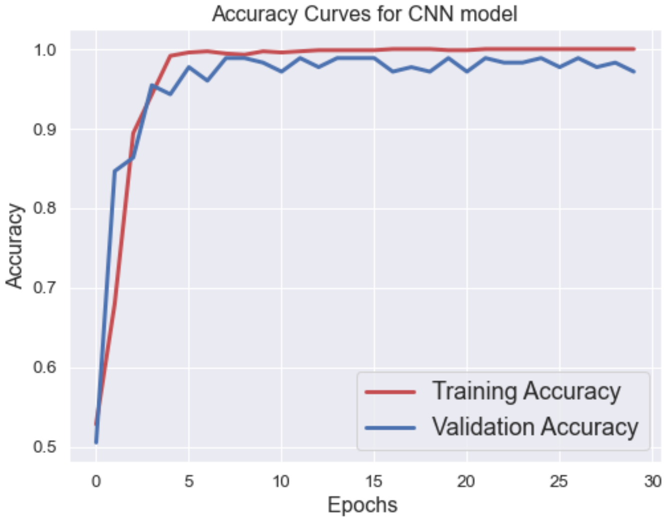


Figure 10: Accuracy Curve

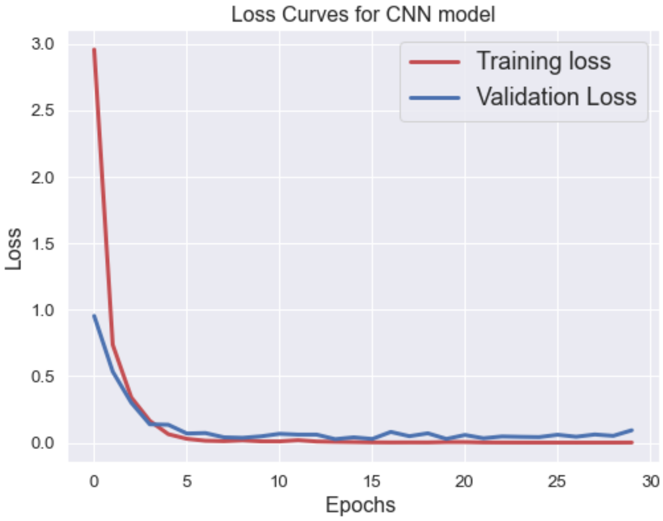


Figure 11: Loss curve

Table 2: Accuracy score of CNN with preprocessed dataset and the TL features

| **Evaluation Parameters** | **Preprocessed raw data** | **Transfer learning features** |
|---|---|---|
| Training Accuracy | 100.00% | 100.00% |
| Test Accuracy | 98.190% | 95.928% |
| Training Loss | 0.001 | 0.00079 |
| Test Loss | 0.044 | 0.103 |
| Precision | 0.983 | 0.962 |
| Recall | 0.982 | 0.959 |
| F1 Score | 0.981 | 0.960 |

For high speed data processing, we apply feature extraction methods to reduce the dimensionality of raw data. First, we apply transfer learning (TL) based VGG16 model for feature extraction and applied the extracted TL features for classification. The performance evaluation scores using CNN classification is shown in Table 2 which shows that 95.928% accuracy is obtained using the extracted TL features. Figure 12 shows few of the predicted results including three false predictions using CNN. Here, "true" indicates the real label and "pred" indicates

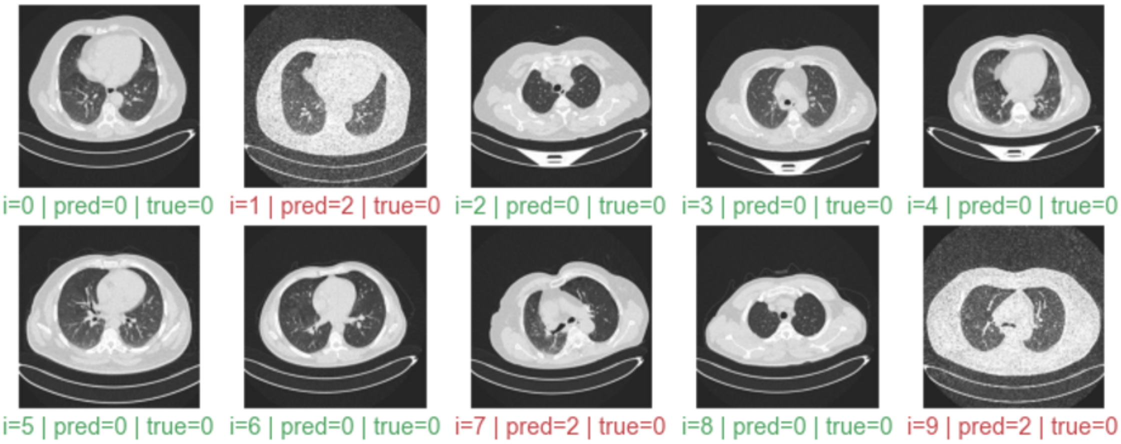


Figure 12: Few of the predicted results using CNN

Table 3: Evaluation Score of Machine learning algorithms with preprocessed raw dataset

| Algorithms | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| DT | 0.876 | 0.869 | 0.871 | 86.88% |
| KNN | 0.969 | 0.968 | 0.967 | 96.83% |
| RF | 0.958 | 0.955 | 0.951 | 95.48% |
| ET | 0.991 | 0.991 | 0.991 | 99.09% |
| XGB | 0.996 | 0.995 | 0.995 | 99.55% |
| SVM | 0.965 | 0.964 | 0.963 | 96.38% |
| LR | 1.000 | 1.000 | 1.000 | 100% |

Table 4: Evaluation Score of machine learning algorithms on the Transfer learning features

| Algorithms | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| DT | 0.969 | 0.968 | 0.969 | 96.83% |
| KNN | 0.969 | 0.968 | 0.967 | 96.83% |
| RF | 0.987 | 0.986 | 0.986 | 98.64% |
| ET | 0.991 | 0.991 | 0.991 | 99.09% |
| XGB | 0.996 | 0.995 | 0.995 | 99.55% |
| SVM | 0.971 | 0.968 | 0.966 | 96.83% |
| LR | 1.000 | 1.000 | 1.000 | 100% |

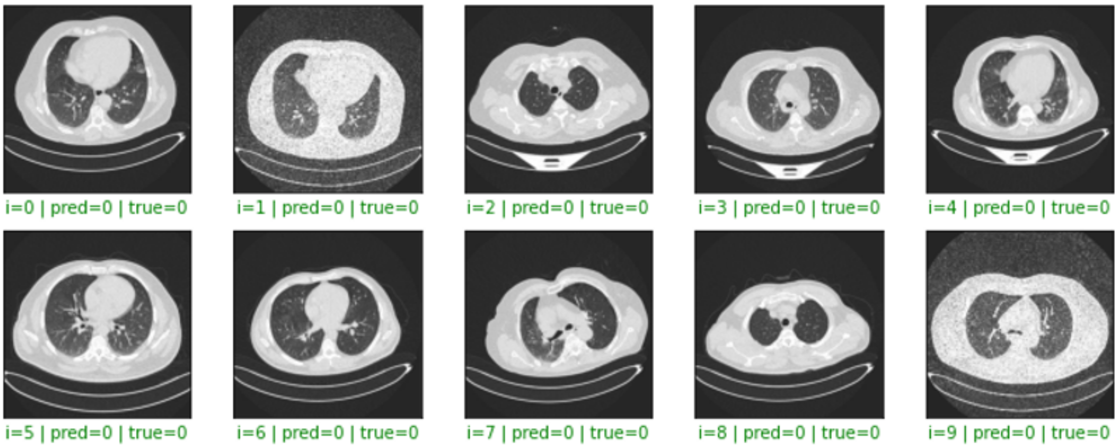


Figure 13: Few of the predicted results using Logistic regression

the predicted label. As we can see that the three false predictions indexed as i=1, i=7 and i=9 detected the benign nodules as normal.

We also apply machine learning based classification algorithms on the preprocessed raw data and TL features. Table 3 and 4 shows the accuracy scores using different machine learning algorithms with preprocessed raw dataset and the TL features, respectively. As we can see that among all machine learning algorithms, the logistic regression (LR) shows 100% accuracy using both preprocessed raw data and extracted TL features. We can also observe that significantly high accuracy is obtained using $ET$ and $XGB$ classifiers. Figure 13 shows few of the prediction results using Logistic regression (LR).

To improve the computational intelligence reducing the computational complexity, we apply the proposed morphological segmentation and feature (MSF) extraction on the preprocessed data. The results of morphological segmentation of malignant, benign and normal nodules are shown in Figs 4, 5 and 6, respectively. The features of the segmented ROI are then extracted us-

Table 5: Evaluation Score of Machine learning algorithms on the Morphological features

| **Algorithms** | **Precision** | **Recall** | **F1 score** | **Accuracy** |
|---|---|---|---|---|
| DT | 0.986 | 0.985 | 0.985 | 98.49% |
| KNN | 0.996 | 0.996 | 0.996 | 99.56% |
| RF | 0.996 | 0.996 | 0.996 | 99.56% |
| ET | 0.996 | 0.996 | 0.996 | 99.56% |
| XGB | 0.996 | 0.996 | 0.996 | 99.56% |
| SVM | 0.997 | 0.997 | 0.997 | 99.71% |
| LR | 0.989 | 0.989 | 0.989 | 98.93% |

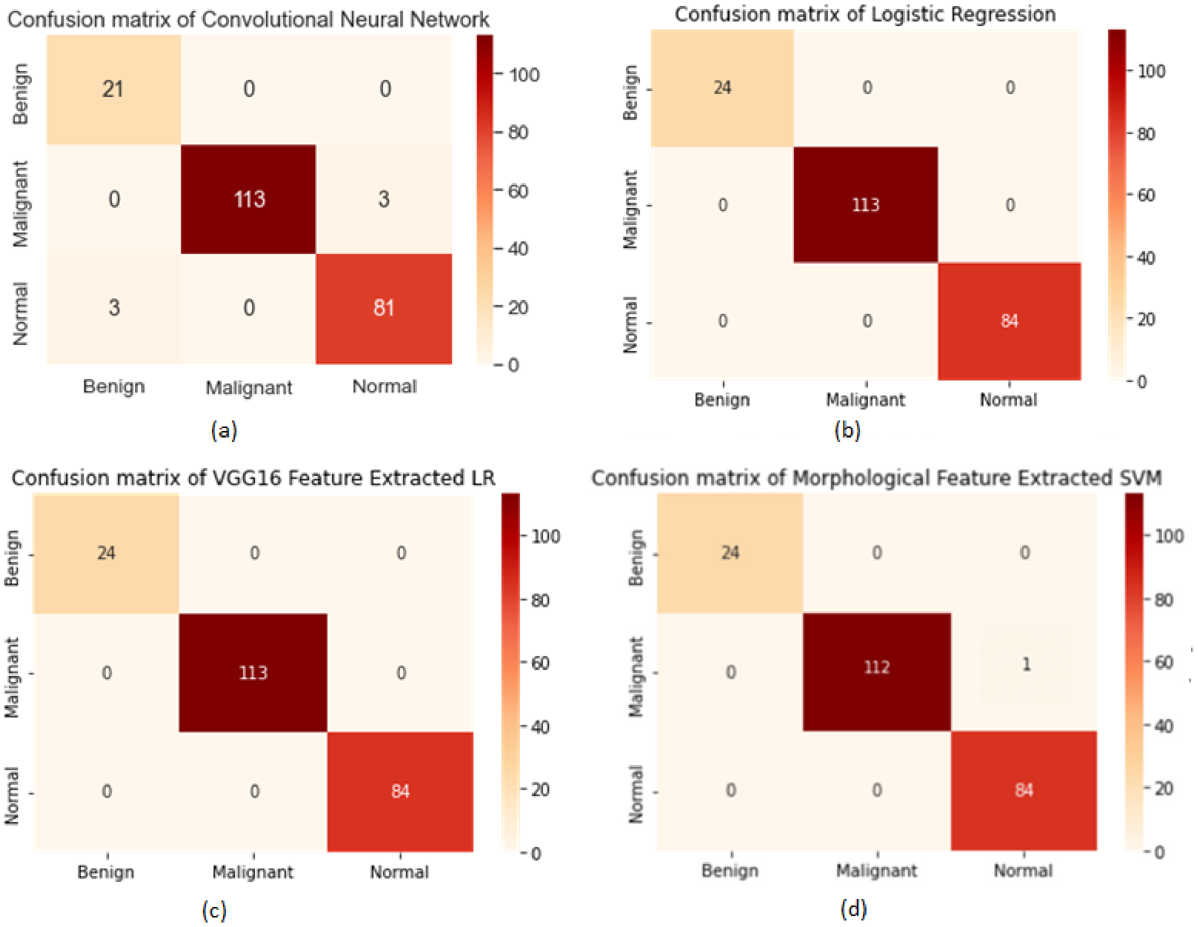


Figure 14: (a) Confusion matrix of CNN using raw data, (b) Confusion matrix of LR using raw data, (c) Confusion matrix of LR using Tranfer learning features, (d) Confusion matrix of SVM using Morphological features.

ing the proposed morphological features extraction methods. The extracted features are applied to the machine learning algorithms for classification. Table 5 shows the accuracy scores of the machine learning approaches on the extracted morphological features. Hence, we can see that all of the machine learning algorithms show significantly high accuracy using the morphological features. Most of the classifiers including KNN, RF, ET, XGB shows 99.56% accuracy where the best accuracy of 99.71% is observed using the SVM classifier.

Figure 14 shows the confusion matrix of CNN classifier and other machine learning based classifiers for the best possible accuracy using both raw data and the extracted features. Figure 14(a) shows the confusion matrix using CNN approach with raw data where it can be seen

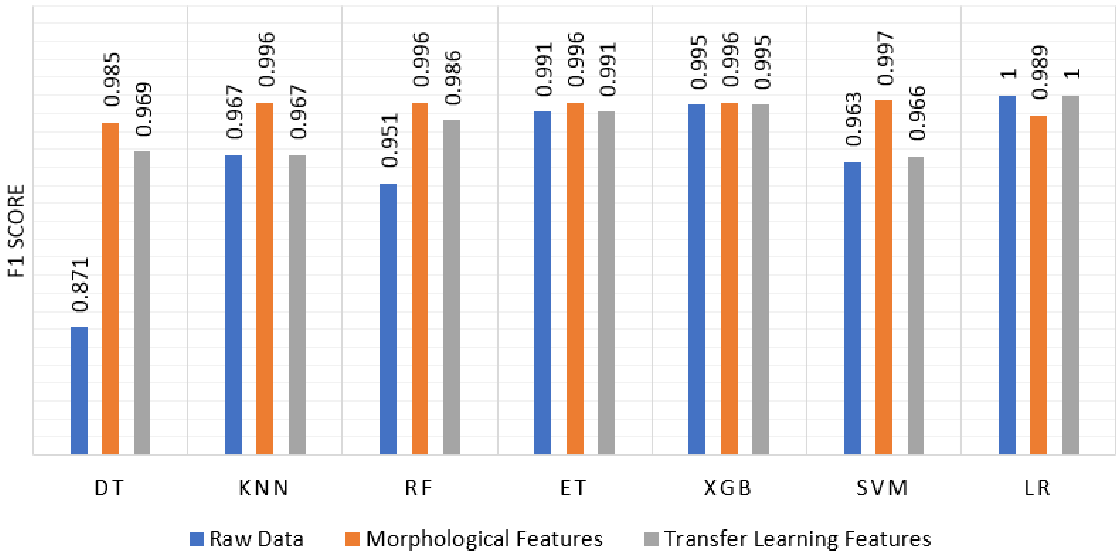


Figure 15: Comparison of the F1 score using different proposed methods

Table 6: Comparison of classification accuracy using different machine learning algorithms

| **Algorithms** | **Preprocessed raw data** | **Transfer learning features** | **Morphological features** |
|---|---|---|---|
| DT | 86.878% | 96.833% | 98.49% |
| KNN | 96.833% | 96.833% | 99.56% |
| RF | 95.475% | 98.643% | 99.56% |
| ET | 99.095% | 99.095% | 99.56% |
| XGB | 99.548% | 99.548% | 99.56% |
| SVM | 99.380% | 96.833% | 99.71% |
| LR | 100% | 100% | 98.93% |

that 3 of the benign nodules are confused as the normal nodules and 3 of the normal nodules are confused as the malignant nodules during classification. Figure 14(b) and (c) shows the confusion matrix using LR approach with raw data and TL features, respectively where we can see that none of the nodules are confused with other types of nodules. Figure 14(d) shows the confusion matrix using SVM with morphological features where we can see that 1 of the malignant type nodules is confused as the normal type nodule.

Figure 15 compares the obtained F1 score of all machine learning classification algorithms using raw data, morphological features and transfer learning features. It can be seen in Fig. 15 that all the proposed algorithms show improved performance with high F1 score. However, it is also observed that all the classification algorithms perform almost equally well with the extracted low dimensional (4×1) morphological features. Table 6 compares the achievable accuracy of all the proposed classification models using machine learning. As we can see in Table 6 that 100% accuracy is achieved by logistic regression (LR) using both raw data and TL features. It is also observed that with the extracted morphological features, all the machine learning algorithms perform equally well with high accuracy. Among those, SVM shows the best accuracy of 99.71%. Table 7 shows the training time (in seconds) comparison for all the machine learning classification algorithms. Here, we can see that raw data requires high processing time for training among which the XGB requires the highest training time of 617.49 *sec.* It is also observed in Table 7 that the training time is reduced slightly by applying transfer learning based feature extraction. However, the morphological feature extraction reduces the training time significantly for all the

Table 7: Training time comparison of all machine learning algorithms

| **Algorithms** | **Raw Data** (sec) | **Morphological features** (sec) | **Transfer Learning features** (sec) |
|---|---|---|---|
| DT | 64.642 | 0.011 | 20.476 |
| KNN | 0.19 | 0.014 | 0.146 |
| RF | 10.973 | 0.339 | 4.955 |
| ET | 8.132 | 0.236 | 6.217 |
| XGB | 617.49 | 0.806 | 300.629 |
| SVM | 73.834 | 0.032 | 44.826 |
| LR | 49.486 | 0.333 | 37.519 |

Table 8: Accuracy comparison to other related works

| **Algorithms** | **Data set** | **Segmentation and Feature extraction** | **Classifier** | **Accuracy** |
|---|---|---|---|---|
| Günaydin et al [4] | CT images | - | Artificial Neural Network (ANN)<br>and Decision Trees (DT) | 82.43%<br>93.24% |
| Punithavathy et al [5] | PET/CT images | Texture and fractal features | SVM with RBF kernel | 98.10% |
| Tao et al [6] | CT images | DenseNet | DenseNet-NSCR | 99.10% |
| Hu et al [8] | CT images | Image segmentation and DenseNet module | Parallel Deep learning with hybrid attention mechanism | 94.61% |
| Moitra et al [9] | PET/CT images | Image segmentation and hybrid feature extraction using MSER-SURF | 1D CNN | 96 ± 3% |
| Boban et al [10] | CT images | Gray level Co-occurance Matrix (GLCM) | MLP,<br>SVM,<br>KNN | 98%,<br>70.45%<br>99.2% |
| Abdullah et al [11] | UCI ML dataset | - | SVM,<br>KNN,<br>CNN | 95.56%,<br>92.11%,<br>88.40% |
| Nanglia et al [14] | CT images | SURF and Generic algorithm (GA) | SVM with FFBPNN | 98.08% |
| Proposed | CT images | - | CNN | 98.19% |
| Proposed | CT images | TL VGG16 features | CNN | 95.92% |
| Proposed | CT images | - | LR | 100% |
| Proposed | CT images | TL VGG16 features | LR | 100% |
| Proposed | CT images | Morphological segmentation and feature extraction | SVM | 99.71% |

classification algorithms by extracting the low dimensional features. Among those, DT requires the lowest training time of 0.011 *sec* to acquire 98.49% accuracy and SVM requires 0.032 *sec* training time to achieve the highest accuracy of 99.71%. Thus, it can be concluded that the low dimensional morphological features can achieve significantly high classification accuracy with low complexity and high computational intelligence. Table 8 compares the accuracy of the proposed works to other related works in the literature. It is apparent that the proposed works outperforms the literature with high computational intelligence and accuracy.

# 7 Conclusion

Early detection and classification of the lung nodules are required to increase the survival rate of the patients. Thus, a system is required with high computational intelligence and low complexity. In this paper, we have proposed lung nodule detection and classification using different preprocessing steps, ROI segmentation, feature extraction and classification methods. First, we have applied different deep learning and machine learning based classification algorithms on the raw CT scan data. It is observed that deep learning based CNN model can acquire 98.19% and machine learning based logistic regression (LR) can acquire 100% accuracy using the preprocessed raw dataset. However, raw data requires high processing time to train the data. Therefore, we have proposed and applied feature extraction methods to extract low dimensional features which requires less processing time for classification. First, we have applied VGG16 model to extract transfer learning (TL) based features of the preprocessed data. Hence, we obtained 95.928% accuracy using CNN and 100% accuracy using LR with the extracted TL features with comparatively low processing time for training. Then, to improve the computational intelligence with low complexity, we have proposed morphological segmentation and feature (MSF) extraction method to extract very low dimensional features. The extracted MSF features are applied for classification using machine learning algorithms. It is observed that most of the machine learning algorithms show significantly high classification accuracy with the extracted morphological features. Due to low dimensional features, it also reduces the training time of all algorithms significantly which increases the processing time of classification. Hence, the highest accuracy of 99.71% is achieved using SVM with the extracted morphological features requiring very low training time of 0.032 seconds. Thus, it may be concluded that our proposed methods of lung nodule detection and classification outperforms the literature with significantly high accuracy with low computational complexity and high intelligence.

# Declarations

- Conflict of interest: The authors declare no conflict of interest.

# References

[1] P. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *IEEE International Conference on Electrical, Computer and Communication Technologies*. IEEE, 2019, pp. 1–4.

[2] K. Manisha, "Differences Between a Malignant and Benign Tumor," http://www.differencebetween.net/science/health/difference-between-benign-and-malignant/, [Online accessed 2022-04-17].

[3] C. Kaushal, S. Bhat, D. Koundal, and A. Singla, "Recent trends in computer assisted diagnosis (CAD) system for breast cancer diagnosis using histopathological images," *Irbm, Elsevier*, vol. 40, no. 4, pp. 211–227, 2019.

[4] Ö. Günaydin, M. Günay, and Ö. Şengel, "Comparison of lung cancer detection algorithms," in *Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science*. IEEE, 2019, pp. 1–4.

[5] K. Punithavathy, S. Poobal, and M. Ramya, "Performance evaluation of machine learning techniques in lung cancer classification from PET/CT images," *FME Transactions*, vol. 47, no. 3, pp. 418–423, 2019.

[6] Z. Tao, H. Bingqiang, L. Huiling, Y. Zaoli, and S. Hongbin, “NSCR-based DenseNet for lung tumor recognition using chest CT image,” *BioMed Research International, Hindawi*, vol. 2020, 2020.

[7] K. Pradhan and P. Chawla, “Medical internet of things using machine learning algorithms for lung cancer detection,” *Journal of Management Analytics, Taylor & Francis*, vol. 7, no. 4, pp. 591–623, 2020.

[8] H. Hu, Q. Li, Y. Zhao, and Y. Zhang, “Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2880–2889, 2020.

[9] D. Moitra and R. K. Mandal, “Classification of non-small cell lung cancer using one-dimensional convolutional neural network,” *Expert Systems with Applications, Elsevier*, vol. 159, p. 113564, 2020.

[10] B. M. Boban and R. K. Megalingam, “Lung diseases classification based on machine learning algorithms and performance evaluation,” in *International Conference on Communication and Signal Processing.* IEEE, 2020, pp. 0315–0320.

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

[12] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, “Prediction and classification of lung cancer using machine learning techniques,” in *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1. IOP Publishing, 2021, p. 012059.

[13] N. Nawreen, U. Hany, and T. Islam, “Lung cancer detection and classification using ct scan image processing,” in *International Conference on Automation, Control and Mechatronics for Industry (ACMI).* IEEE, 2021, pp. 1–6.

[14] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, “A hybrid algorithm for lung cancer classification using SVM and Neural Networks,” *ICT Express, Elsevier*, vol. 7, no. 3, pp. 335–341, 2021.

[15] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[16] M. James, “Hands-on Transfer Learning with Keras and the VGG16 Model,” https://www.learndatasci.com/tutorials/hands-on-transfer-learning-keras/, [Online accessed 2022-07-20].

[17] L. Rokach and O. Maimon, “Decision trees,” in *Data mining and knowledge discovery handbook.* Springer, 2005, pp. 165–192.

[18] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems".* Springer, 2003, pp. 986–996.

[19] L. Breiman, “Random forests,” *Machine learning, Springer*, vol. 45, no. 1, pp. 5–32, 2001.

[20] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning, Springer*, vol. 63, no. 1, pp. 3–42, 2006.

[21] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[22] Y. Zhang, “Support vector machine classification algorithm and its application,” in *International conference on information computing and applications.* Springer, 2012, pp. 179–186.

[23] R. E. Wright, “Logistic regression.” *American Psychological Association*, 1995.

[24] B. Aayush, “Performance Metrics in Machine Learning [Complete Guide],” https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide, [Online accessed 2022-07-21].

[25] W. Rose, “Cross-Entropy Loss and Its Applications in Deep Learning,” [Online accessed 2022-09-02].

[26] “Kaggle Datasets,” https://www.kaggle.com/datasets/antonixx/the-iqothnccd-lung-cancer-dataset, [Online accessed 2022-04-19].