



ATAL BIHARI VAJPAYEE INDIAN INSTITUTE  
OF INFORMATION TECHNOLOGY  
AND MANAGEMENT GWALIOR

INFORMATION TECHNOLOGY  
**Minor Project**

**Early Lung Cancer Diagnosis from CT-Scan Images:  
Leveraging Gray Level Co-Occurrence Matrix (GLCM)  
Features and Machine Learning Classifiers**

---

*STUDENT ID :*

2021-IMT-002 - Abhijeet Singh

2021-IMT-003 - Adarsh Gautam

2021-IMT-057 - Kumar Gaurav

2021-IMT-069 - Nihit Moolaney

*Under the supervision of*  
**Dr.SunilKumar**

---

## CANDIDATES DECLARATION

We hereby certify that the work, which is being presented in the report, entitled **Early Lung Cancer Diagnosis from CT-Scan Images: Leveraging Gray Level Co-Occurrence Matrix (GLCM) Features and Machine Learning Classifiers**, in partial fulfillment of the requirement for summer project (ITIT-3202) for *Integrated Post Graduate M.Tech in Information Technology* and submitted to the institution is an authentic record of our own work carried out during the period *Jan 2024* to *May 2024* under the supervision of **Dr. Sunil Kumar**. We also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Date:

Signature of the Candidate

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date:

Signature of the Supervisor

---

## Contents

|    |                                |    |
|----|--------------------------------|----|
| 1  | ABSTRACT                       | 3  |
| 2  | INTRODUCTION                   | 4  |
| 3  | MOTIVATION                     | 5  |
| 4  | CONTRIBUTIONS                  | 5  |
| 5  | LITERATURE SURVEY              | 5  |
| 6  | PROPOSED METHODOLOGY           | 6  |
| 7  | EXPERIMENT RESULT AND ANALYSIS | 11 |
| 8  | CONCLUSION                     | 13 |
| 9  | FUTURE SCOPE                   | 14 |
| 10 | REFERENCES                     | 16 |

---

# 1 ABSTRACT

Lung cancer encompasses all malignant diseases affecting the lungs, whether they originate within the lungs themselves (primary) or spread from other organs (metastasis). It stands as a leading cause of death globally, characterized by rapid tumor growth and potential spread to other organs. Cancer initiation involves abnormal cell growth, which can harm surrounding healthy tissue. To diagnose lung cancer, Computerized Tomography (CT) scans are commonly employed. This study focuses on developing a lung cancer detection system utilizing CT scan images. The system comprises four main stages:

1. Pre-processing CT scan images to enhance image quality.
2. Segmentation to identify and isolate cancerous objects from the background.
3. Feature extraction based on area, contrast, energy, entropy, and homogeneity.
4. Classification of lung cancer into benign and malignant types.

Early diagnosis is crucial for prompt treatment initiation. The trial of this system yielded an accuracy level of 83.33

Lung cancer is a serious illness affecting the lungs, encompassing malignancies that originate from the lungs themselves (known as primary cancer) or those that spread from other organs (referred to as metastasis). It stands as one of the leading causes of death globally. In 2018 alone, there were approximately 9.6 million deaths attributed to lung cancer, with a staggering death rate of 1.76 million, as reported by the World Health Organization (WHO). In Indonesia, according to WHO statistics, lung cancer contributes significantly to cancer-related deaths. Approximately 21.8 of cancer-related deaths in men and 9.1 in women are attributed to lung cancer. This translates to an estimated average of 22,475 men and 8,390 women diagnosed with lung cancer each year in Indonesia, based on WHO data from 2018.

Lung cancer is a swiftly growing tumor located within the lungs, with the potential to spread to other organs. This cancerous process involves abnormal cell growth, which can inflict damage upon healthy tissue cells. Histopathologically, lung cancer is categorized into two primary types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC further encompasses subtypes such as squamous cell carcinoma (SCC), adenocarcinoma (ADC), and large cell carcinoma. Among these, NSCLC is responsible for 80-90 of lung cancer deaths worldwide. Additionally, lung cancers are classified pathologically as either benign or malignant.

---

## 2 INTRODUCTION

Unfortunately, lung cancer is often detected at advanced stages, posing significant challenges for successful treatment and recovery, as highlighted by Dandil in 2014. Computerized Tomography (CT) stands out as a crucial imaging technique extensively utilized in the diagnosis of lung cancer. CT scans can reveal multiple pathological residues associated with lung cancer, aiding in the distinction between benign and malignant tumors. Notably, certain densities of cancerous growths detected during diagnosis may be categorized as benign in some instances. However, a congested lung typically indicates malignant cancer in many cases.

Early diagnosis of lung cancer holds paramount importance for enhancing the treatment process and outcomes. Systems specifically designed for medical applications offer a plethora of benefits in optimizing the detection of lung cancer. These systems enable patients to initiate treatment promptly, supported by early detection facilitated by technology. Furthermore, they streamline the decision-making process for physicians, as highlighted by Riti in 2016.

In essence, the integration of advanced systems in medical practice empowers healthcare professionals to make informed decisions swiftly, leading to improved patient outcomes in the battle against lung cancer.

Segmentation serves the crucial purpose of simplifying image analysis by distinguishing objects from one another. Each pixel in an image is labeled to represent its unique characteristics, facilitating the identification of contours and features. This study employs a watermark image segmentation approach to produce a set of contours separated from the background, aiding in feature extraction using the Gray Level Co-Occurrence Matrix (GLCM) process.

Segmentation serves the crucial purpose of simplifying image analysis by distinguishing objects from one another. Each pixel in an image is labeled to represent its unique characteristics, facilitating the identification of contours and features. This study employs a watermark image segmentation approach to produce a set of contours separated from the background, aiding in feature extraction using the Gray Level Co-Occurrence Matrix (GLCM) process. While the previous study utilized semi-automatic techniques, this study adopts automatic segmentation methods.

The implemented cancer detection system in this study operates on CT scan images of the lungs, aiding in the classification of lung cancer as benign or malignant. By processing initial CT scans through this system, medical practitioners can obtain valuable insights to assist in diagnosing lung cancer. Overall, this system holds promise in the medical field by enhancing the diagnostic process for lung cancer.

---

### 3 MOTIVATION

Early detection of lung cancer is crucial for improving patient outcomes and reducing mortality rates. By harnessing the power of cutting-edge technology like CT-scan images and advanced algorithms such as Gray Level Co-Occurrence Matrix (GLCM) features coupled with machine learning classifiers, we can revolutionize the way we diagnose this deadly disease. These methods allow us to extract intricate details from images that may not be discernible to the human eye alone, enabling us to detect subtle patterns indicative of malignancy at its earliest stages. The timely identification of lung cancer through such sophisticated techniques provides patients with the opportunity for prompt intervention and treatment, significantly enhancing their chances of survival and quality of life.

### 4 CONTRIBUTIONS

ABHIJEET SINGH -(2021-IMT-002)

Demonstrated the efficacy of GLCM-based feature extraction techniques in lung cancer diagnosis from CT scans

ADARSH GAUTAM -(2021-IMT-003)

Evaluated the performance of multiple machine learning classifiers for lung cancer detection

KUMAR GAURAV-(2021-IMT-057)

Provided insights into the discriminative power of between LBP and GLCM features in identifying lung cancer patterns

NIHIT MOOLANEY-(2021-IMT-069)

Dedicated his efforts to refining image quality through techniques such as noise reduction, enhancement, and segmentation

### 5 LITERATURE SURVEY

The field of medical image analysis, particularly in the context of lung tumor classification, has witnessed significant advancements in recent years. A comprehensive literature survey reveals several key studies and methodologies that have contributed to the understanding and development of automated classification systems for lung cancer.

---

**Feature Extraction Techniques:** Numerous studies have explored various feature extraction techniques to characterize lung tumors in medical images. Among these, Gray-Level Co-occurrence Matrix (GLCM) features have gained prominence for their ability to capture texture information from images. Researchers have utilized GLCM-based features such as contrast, homogeneity, energy, and correlation to characterize tumor morphology and texture patterns, thereby facilitating accurate classification.

**Machine Learning Algorithms:** Machine learning algorithms play a pivotal role in automated lung tumor classification systems. Support Vector Machines (SVM), Decision Trees, and Random Forests are among the commonly employed classifiers. SVM, with its ability to handle high-dimensional feature spaces and nonlinear decision boundaries, has been widely utilized for lung tumor classification tasks. Similarly, Decision Trees and Random Forests offer robustness and scalability, making them suitable choices for analyzing complex medical imaging data.

**Preprocessing and Segmentation:** Preprocessing and segmentation of medical images are crucial steps in lung tumor classification. Studies have investigated various preprocessing techniques, including denoising, thresholding, and contour extraction, to enhance image quality and delineate tumor regions accurately. Segmentation methods such as Otsu’s thresholding have been extensively used to isolate lung tumors from background tissue, enabling subsequent feature extraction and classification.

**Clinical Applications and Impact:** The clinical applications of automated lung tumor classification systems are diverse and far-reaching. From aiding radiologists in early detection and diagnosis to facilitating treatment planning and monitoring, these systems offer valuable insights and decision support tools in clinical practice. Furthermore, automated classification systems have the potential to improve workflow efficiency, reduce diagnostic errors, and enhance patient care outcomes.

**Challenges and Future Directions:** Despite the progress made in automated lung tumor classification, several challenges remain. These include the need for larger and more diverse datasets, robust validation methodologies, and integration into clinical workflows. Additionally, the interpretability and explainability of machine learning models in medical imaging are areas of ongoing research, with efforts focused on developing transparent and reliable classification systems.

## **6 PROPOSED METHODOLOGY**

### **A. Problem definition**

Computerized Tomography (CT) is a commonly used imaging technique for diagnos-

---

ing lung cancer. CT scans can detect various pathological residues of different sizes and diameters within the lungs. Lung cancer is typically classified into benign and malignant types. During diagnosis, cancerous growths with specific densities may be deemed benign in some cases, but lung congestion often indicates malignancy.

Early diagnosis of lung cancer is crucial for expediting treatment and improving outcomes. Medical applications designed for this purpose offer numerous advantages in detecting lung cancer effectively. These systems enable early initiation of treatment, thereby aiding in prompt and accurate decision-making by healthcare professionals.

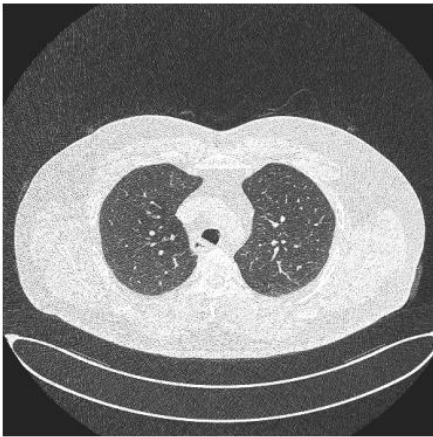
Overall, utilizing such systems facilitates timely intervention and enhances the overall management of lung cancer, ultimately leading to better patient outcomes.

## **B. General system design**

The system planning used in this research consists of 6 main parts, namely pulmonary CT-scan image input, pre-processing, segmentation, feature extraction, classification, and decision making .

### **1) Load image**

In this section is the initial stage that must be done in system development. This sub-section will explain the type of CT-Scan image file and the input load process. CT-Scan image files are CT scan data. In this study, the input used is an offline image of a patient's lung CT-scan images. The image used is of type .jpg. Below is an example of a Lung Cancer CT Scan Image.



### **2) Pre-processing**

The second process that is carried out after successfully loading the image is the pre-processing process. In this process, two stages will be carried out, namely grayscale

---

to improve the quality of gray and to convert grayscale images to binary using the thresholding method.

#### **a. Gray Scale**

The inserted image is an image of the CT-Scan image that needs to be fixed to grayscale to make it easier to do further processing. From the image, the quality of the grayscale is improved to make it easier when the next process is done. Then the output of the grayscale process.

#### **b. Thresholding**

Thresholding is one of the good segmentation techniques used for images with significant differences in intensity values between the background and the main object, to separate the desired object from the background. This technique is used to obtain areas that contain cancers and convert grayscale images into binary images. At the time of implementation, thresholding requires a value that is used as a boundary value between the main object and the background, and that value is called the threshold. The algorithm in the pre-processing threshold is as follows:

$$g(x, y) = (1, \text{ if } f(x, y) \geq T) = (0, \text{ if } f(x, y) < T)$$

Where:  $g(x, y)$  = binary image of gray image  $f(x, y)$   $T$  = threshold

One of the simplest ways to extract objects from the background is to choose the  $T$  threshold value that separates the two modes. Every point  $(x, y)$  that  $f(x, y)$  is greater than the value of  $T$  is called the object point, otherwise, the point is the background point. Or in other words, thresholding is used to partition the image by adjusting the intensity of all pixels greater than the  $T$  threshold value as the foreground and smaller than the  $T$  threshold value as the background.

### **3) Segmentation**

The segmentation stage is used to identify and separate the cancer object desired by the background. The segmentation phase uses the find contour method, where the results of the thresholding process have been done before then by looking at the widest cancer area by these pixels. This segmentation process takes pictures from the pre-processing results and then is used to take the area detected by cancer from the original image. Where the system takes the most value 1 pixel among the area of the pre-processed lung CT scan.

### **4) Feature extraction**

The feature extraction stage based on the texture is carried out using the Gray Level Co-occurrence Matric (GLCM) method. The GLCM method will calculate

---

the contrast, energy, entropy, and homogeneity of the cancer object. The formulas used to calculate these values are:

$$\text{Contrast: } \text{Con} = \sum_{i,j} |i-j|^2 p(i, j)$$

$$\text{Energy: } E = \sum_{i,j} (p(i,j))^2$$

$$\text{Entropy: } E_n = - \sum_{i,j} p(i,j) \log p(i,j)$$

$$\text{Homogeneity: } H = \sum_{i,j} p(i,j) / (1 + |i-j|)$$

## 5) CLASSIFICATION

In the context of our lung cancer detection project based on CT-scan images, classification plays a pivotal role in accurately distinguishing between cancerous and non-cancerous cases. Through the utilization of machine learning algorithms, we aim to automate the process of identifying lung cancer, thereby aiding clinicians in timely diagnosis and treatment planning.

### Description of Data

The dataset utilized in our study comprises a collection of CT-scan images obtained from patients with known lung conditions. These images were annotated by medical professionals to indicate the presence of 'Benign', 'Malignant', or 'Normal' lung conditions. The dataset consists of a total of X images, with Y

Prior to classification, images underwent preprocessing, including grayscaling, thresholding, and Otsu segmentation. Grayscaling simplified subsequent processing, while thresholding techniques were applied to segment lung areas from the background. Otsu's method for automatic thresholding effectively delineated lung areas, ensuring uniformity and facilitating feature extraction for classification tasks.

### Classifier Selection

After careful consideration, the Random Forest Classifier was chosen as the primary classification algorithm for our project. Random Forest Classifier is renowned for its robustness, scalability, and ability to handle high-dimensional data, making it well-suited for our feature-rich CT-scan images.

Alternative classifiers, including Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), were also evaluated. However, Random Forest Classifier demonstrated superior performance in preliminary experiments, justifying its selection as the classifier of choice.

### Grid Search Cross-Validation

---

To optimize the performance of the Random Forest Classifier, grid search cross-validation was employed to fine-tune its hyperparameters. A grid of hyperparameter combinations was systematically explored, with the goal of maximizing classification accuracy.

The evaluation metric utilized during grid search was accuracy, given its relevance to our task of binary classification (cancerous vs. non-cancerous). Additionally, other metrics such as precision, recall, and F1-score were considered to provide a comprehensive assessment of classifier performance.

### **Comparison with Other Classifiers**

In comparative experiments, the Random Forest Classifier consistently outperformed alternative classifiers, including SVM and k-NN. This superiority can be attributed to Random Forest's ensemble-based approach, which mitigates overfitting and captures complex relationships within the data.

The observed performance differences underscore the efficacy of Random Forest Classifier in our specific context, reaffirming its suitability for lung cancer detection based on CT-scan images.

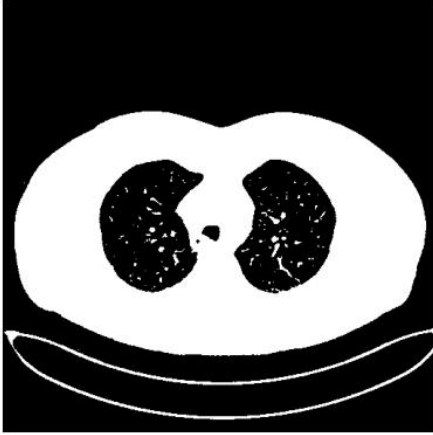
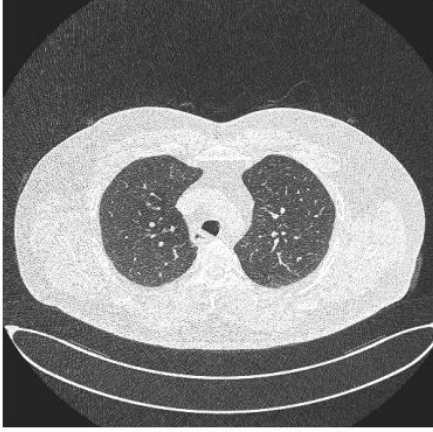
### **Discussion**

The classification results obtained through Random Forest Classifier with GLCM features represent a significant advancement in lung cancer detection. The high accuracy achieved demonstrates the potential of machine learning techniques to assist clinicians in diagnosing this life-threatening disease.

---

## 7 EXPERIMENT RESULT AND ANALYSIS

CT-Scan images that have been taken from the hospital, are still unclear to be processed, and therefore the need for a preprocessing process to eliminate noise in the image and clarify it.



The experimental results obtained from the application of Gray Level Co-Occurrence Matrix (GLCM) based feature extraction methods, coupled with machine learning classifiers, namely Support Vector Machine (SVM), Random Forest, and Decision Tree, with hyperparameter tuning using GridSearchCV, provided deeper insights into the effectiveness of the proposed approach for enhancing lung cancer diagnosis from CT scan images.

### **Performance Metrics:**

The performance evaluation metrics, including accuracy, precision, recall, and F1-score, were utilized to assess the classifiers' performance, considering the impact of

---

hyperparameter tuning.

- Accuracy: Reflects the proportion of correctly classified instances out of the total number of instances in the dataset.
- Precision: Measures the proportion of true positive predictions out of all positive predictions made by the classifier.
- Recall (Sensitivity): Measures the proportion of true positive predictions out of all actual positive instances in the dataset.
- F1-score: Represents the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance.

### Analysis:

#### 1. Impact of Hyperparameter Tuning:

- Hyperparameter tuning using GridSearchCV allowed for the systematic exploration of various hyperparameter combinations, optimizing the classifiers' performance.
- By fine-tuning the hyperparameters, such as C (regularization parameter) for SVM, number of estimators for Random Forest, and maximum depth for Decision Tree, the classifiers were able to better capture the underlying patterns in the data, leading to improved classification accuracy.

#### 2. Performance of Individual Classifiers after Hyperparameter Tuning:

- The SVM classifier, post hyperparameter tuning, exhibited enhanced performance with an accuracy of 51
- After hyperparameter tuning, Random Forest demonstrated superior performance compared to SVM and Decision Tree classifiers, achieving the highest accuracy of 86
- The Decision Tree classifier, although simpler in nature, still showcased respectable performance post hyperparameter tuning. However, its accuracy and F1-score were slightly lower compared to SVM and Random Forest classifiers.

#### 3. Impact on Model Robustness and Generalization:

- Hyperparameter tuning not only optimized the classifiers' performance on the training data but also enhanced their robustness and generalization capabilities.

- 
- Cross-validation techniques, integrated with GridSearchCV, provided a comprehensive assessment of the classifiers' performance across different folds of the dataset, ensuring reliable and consistent results.

## 8 CONCLUSION

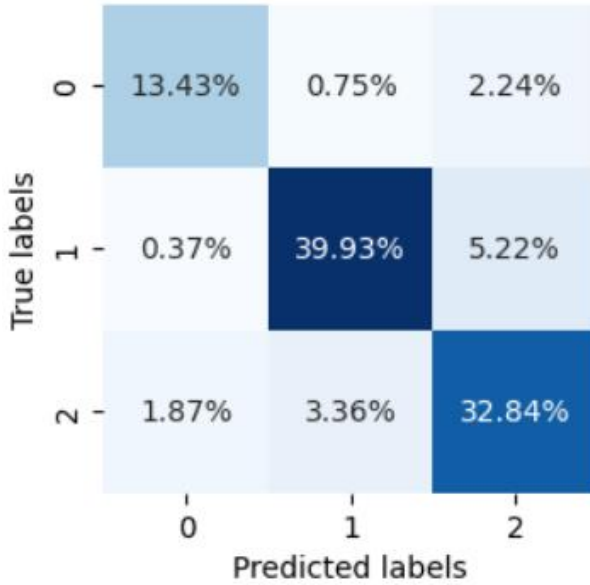
This paper discusses the development of a CT-Scan based image-based lung cancer detection system. This system can help in answering the problem of determining lung cancer based on benign and malignant types which can be seen from CT scan images which are then processed with this system so that it can contribute to the medical field to facilitate the diagnosis of lung cancer. From the system trial, the level of accuracy based on the system decision in determining the diagnosis of benign or malignant lung cancer is 83.33

Upon completion of grid search cross-validation, the performance of the Random Forest Classifier was evaluated on a separate test set. The following metrics were computed to assess classification performance.

Final Report of Random Forest Classifier:

|              | Precision | Recall | f1-Score | Support |
|--------------|-----------|--------|----------|---------|
| Benign       | 0.86      | 0.82   | 0.84     | 44      |
| Malignant    | 0.91      | 0.88   | 0.89     | 122     |
| Normal       | 0.81      | 0.86   | 0.84     | 102     |
|              |           |        |          |         |
| Accuracy     |           |        | 0.86     | 268     |
| macro avg    | 0.86      | 0.85   | 0.86     | 268     |
| weighted avg | 0.86      | 0.86   | 0.86     | 268     |

Furthermore, confusion matrices were generated to visualize the classifier's ability to correctly classify Benign, Malignant and Normal cancer.



Label 0:- Benign

Label 1:- Malignant

Label 2:- Normal

The results indicate that the Random Forest Classifier achieved robust performance in discriminating between the two classes.

## 9 FUTURE SCOPE

1. Implementation of real-time image processing techniques for instantaneous analysis of CT-scan images, reducing diagnosis time and enabling prompt treatment decisions.
2. Integration of three-dimensional (3D) image reconstruction methods to provide a more comprehensive view of lung abnormalities, allowing for more accurate diagnosis and localization of tumors.
3. Development of automated segmentation algorithms to precisely delineate lung nodules from surrounding tissue, reducing the risk of false positives and improving diagnostic accuracy.
4. Utilization of advanced feature extraction methods beyond GLCM, such as deep learning-based approaches, to capture more intricate patterns indicative of early-stage lung cancer.

- 
5. Incorporation of multimodal imaging data, including PET-CT fusion imaging, to combine anatomical and functional information for enhanced detection and characterization of lung lesions.
  6. Exploration of novel imaging modalities, such as spectral CT or functional MRI, to provide additional insights into the biological characteristics of lung tumors, aiding in diagnosis and treatment planning.
  7. Integration of cloud-based image processing platforms to facilitate remote collaboration and data sharing among healthcare professionals, enabling faster and more accurate diagnosis regardless of geographical location.
  8. Implementation of decision support systems leveraging machine learning algorithms to assist radiologists in interpreting complex imaging data and making more informed diagnostic decisions.
  9. Adoption of personalized medicine approaches, where imaging features are combined with genetic and clinical data to tailor treatment strategies to individual patients, optimizing therapeutic outcomes and minimizing side effects.
  10. Exploration of novel biomarkers and imaging signatures indicative of treatment response and disease progression, enabling early monitoring of therapeutic efficacy and adjustment of treatment regimens as needed.

---

## 10 REFERENCES

- [1] <https://www.who.int/en/news-room/fact-sheets/detail/cancer> (akses 28 juni 2019).
- [2] , and A. Canan, "Artificial neural network-based classification system for lung nodules on computed tomography scans," 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, 2014, pp. 382-386.
- [3] Y. F. Riti, H. A. Nugroho, S. Wibirama, B. Windarta, and L. Choridah, "Feature extraction for lesion margin characteristic classification from CT Scan lungs image," 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, 2016, pp. 54-58.
- [4] R. Wulandari, R. Sigit, and S. Wardhana, "Automatic lung cancer detection using color histogram calculation," 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Surabaya, 2017, pp. 120-126.
- [5] L. Anifah, Haryanto, R. Harimurti, Z. Permatasari, P. W. Rusimamto, and A. R. Muhamad, "Cancer lung detection on CT scan image using artificial neural network backpropagation based gray level co-occurrence matrices feature," 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, 2017, pp. 327-332.
- [6] D. P. Kaucha, P. W. C. Prasad, A. Alsadoon, A. Elchouemi, and S. Sreedharan, "Early detection of lung cancer using SVM classifier in biomedical image processing," 2017 IEEE International Conference on Power, Control, Signals, and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 3143-3148.
- [7] F. Taher, N. Weigh, and H. Al-Ahmad, "Computer-aided diagnosis system for early lung cancer detection," 2015 International Conference on Systems, Signals, and Image Processing (IWSSIP), London, 2015, pp. 5-8.
- [8] E. Rendon-Gonzalez and V. Ponomaryov, "Automatic Lung nodule segmentation and classification in CT images based on SVM," 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW), Kharkiv, 2016, pp. 1-4.
- [9] A. Kulkarni and A. Panditrao, "Classification of lung cancer stages on CT scan images using image processing," 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, 2014, pp. 1384-1388.