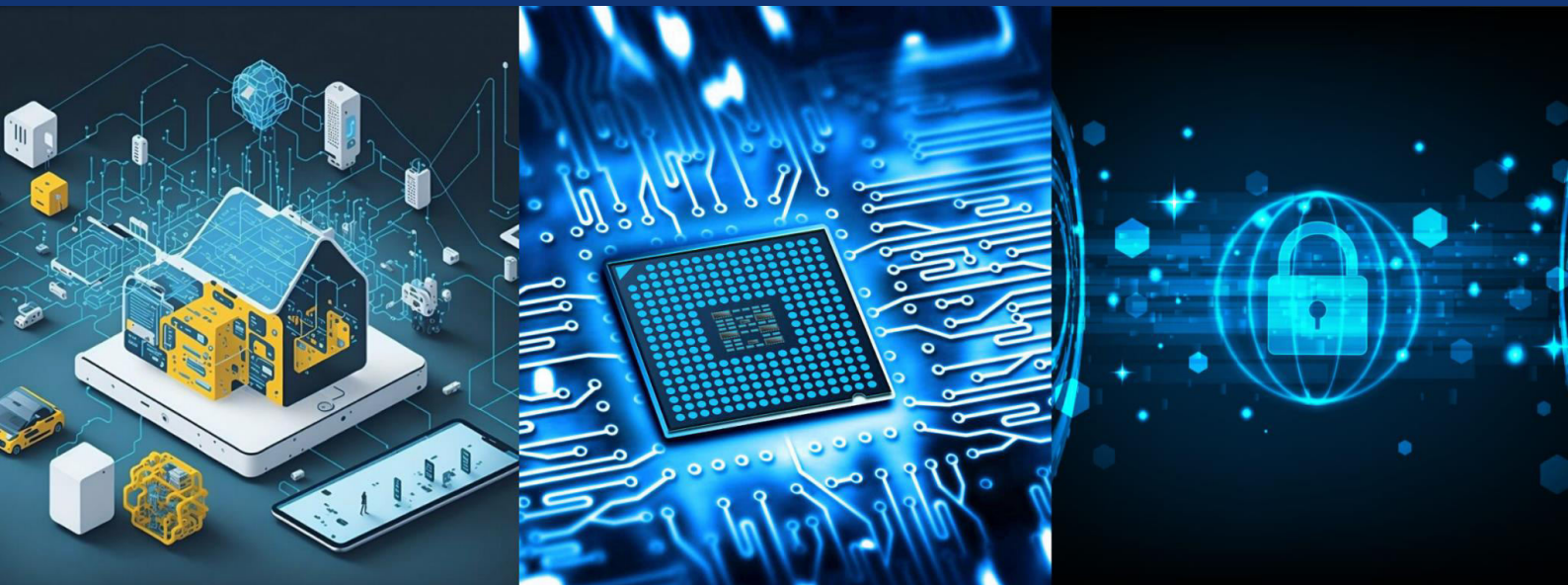
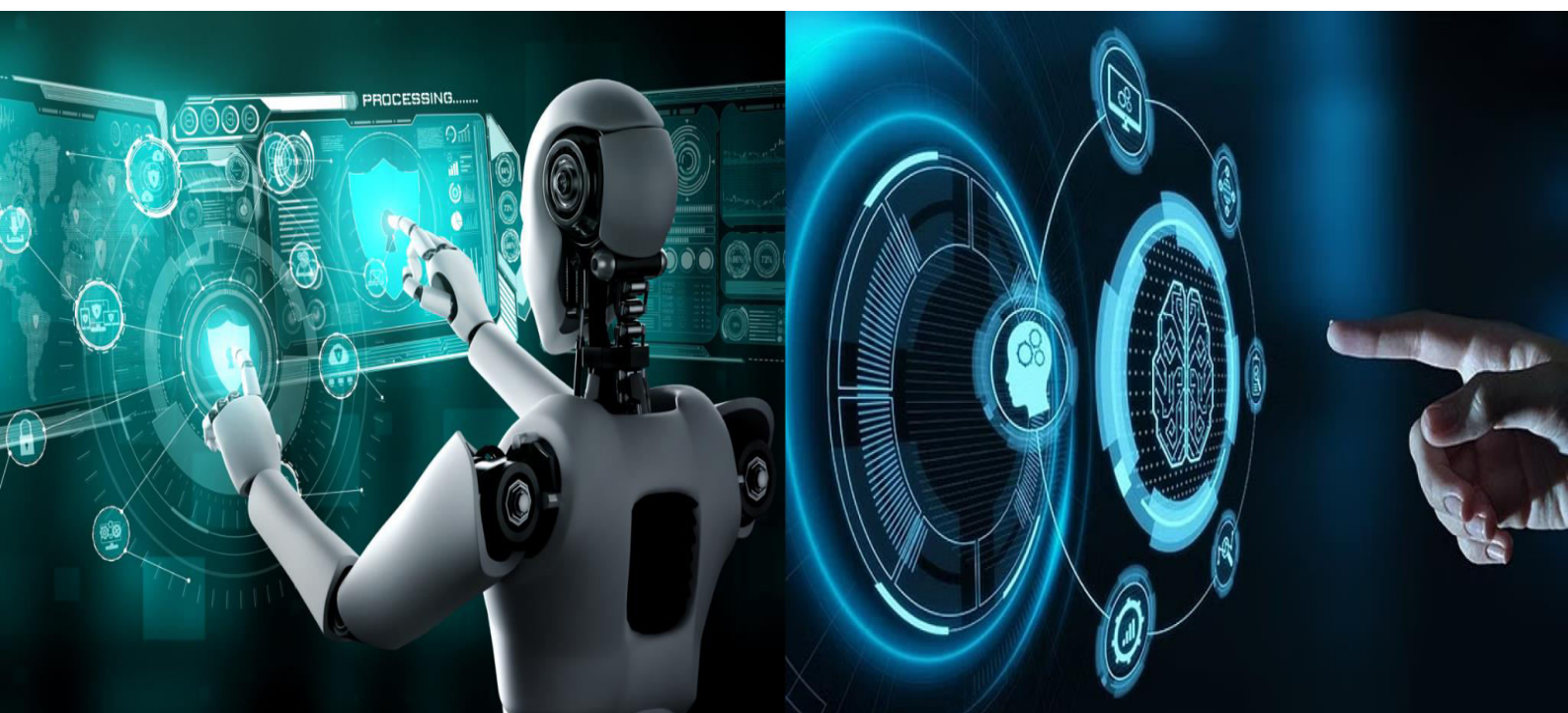




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Medical Image Processing and Machine Learning Based On Lung Cancer Detection Using CT Scan Images

Safiya Haala, Kruthika VT

Student, Department of MCA, CMR University, Bhuvanagiri, OMBR Layout, Bengaluru, India

Assistant Professor, SSCS, OMBR, CMR University, Karnataka, India

ABSTRACT: Lung cancer is one of the major causes of cancer-related deaths, and early detection is important for better treatment outcomes. Computed Tomography (CT) scans provide detailed information about lung structures and are commonly used to identify abnormal tissues. However, manually analyzing a large number of medical images can be time-consuming and may lead to variations in the diagnostic process.

This paper presents a medical image processing and machine learning-based framework for detecting lung cancer from CT scan images. The proposed approach combines image preprocessing, frequency-domain analysis, feature extraction, and supervised machine learning. CT images from a publicly available Kaggle dataset are first resized and converted to grayscale. Median filtering is then applied to reduce noise while preserving important image structures. Fast Fourier Transform (FFT) is used to convert the processed images into the frequency domain and extract frequency-based features.

The extracted features are classified using four machine learning algorithms: **Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN)**. An 80:20 training-testing split is used for evaluation. The models are compared using **Accuracy, Precision, Recall, F1-Score, ROC curves, and Area Under the Curve (AUC)**. Confusion matrices and performance comparisons are also used to examine the classification results.

The experimental results demonstrate how conventional image processing and machine learning techniques can be combined for computer-aided lung cancer analysis. The performance of the selected classifiers is compared to determine their effectiveness on the given CT scan dataset.

KEYWORDS: Lung Cancer Detection, Medical Image Processing, CT Scan, Median Filtering, Fast Fourier Transform, Machine Learning, SVM, Random Forest, KNN, Decision Tree.

I. INTRODUCTION

Lung cancer is one of the major diseases affecting the respiratory system, and early detection is important for effective diagnosis and treatment. Computed Tomography (CT) scans provide detailed information about the lungs and can help identify abnormal regions. However, manually analyzing large numbers of medical images can be time-consuming and may be affected by image quality, noise, and variations in interpretation. Medical image processing and machine learning can help automate and support this analysis.

A typical medical image classification system involves several stages, including image acquisition, preprocessing, feature extraction, and classification. In this work, CT scan images are resized to 128×128 pixels and converted to grayscale. Median Filtering is applied to reduce noise while preserving important image structures. Fast Fourier Transform (FFT) is then used to obtain the frequency-domain representation of the processed images, and the resulting magnitude spectrum is used as the feature representation.

The extracted features are classified using four machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN). Their performance is evaluated using Accuracy, Precision, Recall, F1-Score, and AUC. The dataset used in this study contains four classes: Adenocarcinoma, Large Cell Carcinoma, Normal, and Squamous Cell Carcinoma.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The main contributions of this work are the application of Median Filtering for noise reduction, FFT for frequency-domain feature extraction, machine learning-based classification, and comparative performance evaluation of the selected models.

The remainder of this paper is organized as follows: Section II discusses related work; Section III describes the existing system and problem statement; Section IV presents the proposed methodology; Section V describes the implementation and experimental setup; Section VI presents the experimental results; Section VII discusses the prediction pipeline and results; Section VIII discusses the overall findings and limitations; and Section IX presents the conclusion and future scope.

II. RELATED WORK

The rapid development of medical image processing, artificial intelligence, and machine learning has resulted in increased research on automated lung cancer detection and computer-aided diagnosis systems. Researchers have developed various approaches for analyzing Computed Tomography (CT) scan images, with particular focus on image preprocessing, segmentation, feature extraction, classification, and performance evaluation. These approaches aim to assist medical professionals by improving the efficiency and reliability of lung cancer detection.

Several studies have investigated image preprocessing techniques such as Median Filtering, Gaussian Filtering, Histogram Equalization, and Fast Fourier Transformation (FFT) for improving the quality of CT scan images before classification. These techniques help reduce unwanted noise, enhance image information, and preserve important structural details within medical images. Proper preprocessing is an important stage because the quality of the input image can directly affect the features extracted by subsequent classification algorithms.

Traditional machine learning algorithms have also been widely investigated for lung cancer classification. Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) have been applied because of their relatively simple implementation and ability to classify extracted image features. These approaches generally depend on suitable preprocessing and feature extraction techniques to provide meaningful information to the classification models. The reviewed literature indicates that classical machine learning remains useful, particularly when computational resources and dataset sizes are limited.

More recent research has focused on deep learning techniques for automated lung cancer detection. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), YOLO-based models, and hybrid architectures have been investigated for automatically learning discriminative features from medical images. Other approaches have explored U-Net-based segmentation, attention mechanisms, transfer learning, and feature fusion to improve the representation and classification of lung abnormalities. These methods can provide strong performance on sufficiently large datasets, but they generally require greater computational resources and more complex training procedures.

Researchers have also investigated explainable artificial intelligence (XAI) and attention-based approaches to improve the interpretability of automated medical diagnosis systems. XAI techniques can provide additional information about the basis of model predictions, while attention mechanisms allow deep learning models to focus on important regions of CT scan images. Other studies have explored ensemble learning and transfer learning to improve classification robustness and reduce training requirements. These developments demonstrate the increasing importance of both performance and interpretability in computer-aided lung cancer diagnosis.

Many existing approaches focus primarily on advanced deep learning architectures, which can require significant computational resources and large amounts of training data. In contrast, the proposed work investigates a comparatively lightweight medical image processing pipeline that combines Median Filtering and Fast Fourier Transformation with conventional machine learning algorithms. The system compares four classifiers—Support Vector Machine (SVM), Random Forest, Decision Tree, and K-Nearest Neighbors (KNN)—using a common feature representation and evaluates their performance using Accuracy, Precision, Recall, F1-Score, ROC Curve, and Area Under the Curve (AUC).

As the above literature demonstrates, lung cancer detection has progressed from conventional image processing and handcrafted feature-based classification toward advanced deep learning, transformer-based, attention-based, and explainable AI approaches. However, there remains a need for comparatively simple and computationally manageable



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

frameworks that combine effective preprocessing with multiple machine learning classifiers and provide a direct performance comparison. The present study addresses this requirement by integrating Median Filtering, FFT-based feature extraction, and four machine learning classification algorithms into a unified lung cancer detection framework.

III. EXISTING SYSTEM

The current lung cancer detection systems use medical imaging techniques to analyze CT scan images and identify abnormal lung tissues. Traditional diagnosis mainly depends on radiologists manually examining CT scan images. Although medical experts can identify important abnormalities from CT images, manually analyzing a large number of medical images can be time-consuming and may be affected by variations in image quality, noise, and human error. Computer-Aided Diagnosis (CAD) systems have therefore been developed to assist medical professionals by applying image processing and machine learning techniques to medical images.

A. Traditional Medical Image Processing and Classification

Traditional lung cancer detection systems generally perform several image processing operations before classification. CT scan images are first acquired from medical datasets and then subjected to preprocessing techniques to improve image quality. Common preprocessing methods include Median Filtering, Gaussian Filtering, Histogram Equalization, and other enhancement techniques. These methods are used to reduce noise, improve image quality, and preserve important structural information in the CT scan images.

After preprocessing, relevant image features are extracted and provided to machine learning classifiers. Traditional classifiers such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) have been used for classification of medical images. These approaches can provide useful results when appropriate features are extracted from the input images. However, their performance depends strongly on the quality of preprocessing and feature representation.

B. Limitations of Existing Systems

Existing medical image processing systems have several limitations. Medical CT scan images may contain noise, unwanted artifacts, variations in intensity, and differences in image quality, which can make accurate classification difficult. If preprocessing is not performed effectively, important features may be lost or irrelevant information may be provided to the classification algorithms.

Another limitation is the dependence of traditional machine learning models on manually designed or extracted features. Unlike modern deep learning approaches, conventional machine learning algorithms generally do not automatically learn complex hierarchical image representations. Therefore, the effectiveness of the classification system depends on the selected preprocessing and feature extraction techniques.

Recent deep learning-based systems, including Convolutional Neural Networks (CNNs), Vision Transformers, attention-based models, and hybrid architectures, can automatically learn complex image features and achieve strong performance on large datasets. However, these approaches generally require larger datasets, greater computational resources, and more complex training procedures.

Therefore, there is a requirement for a comparatively simple and computationally manageable framework that combines effective image preprocessing, frequency-domain feature extraction, and multiple machine learning algorithms for comparative evaluation. The proposed system addresses these limitations by integrating Median Filtering, Fast Fourier Transformation (FFT), feature extraction, and four machine learning classifiers into a unified lung cancer detection pipeline.

IV. PROPOSED SYSTEM

The proposed system is designed to improve the efficiency of lung cancer detection from CT scan images by integrating image preprocessing, frequency-domain analysis, feature extraction, and machine learning-based classification. The framework uses Median Filtering to reduce image noise and improve image quality, followed by Fast Fourier Transform (FFT) to obtain frequency-domain information from the processed images. The extracted features are then provided to multiple machine learning classifiers, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(DT), and K-Nearest Neighbors (KNN). The overall workflow provides a structured approach for transforming raw CT scan images into a final predicted lung cancer class.

A. Dataset and Image Preprocessing

The proposed framework begins with a CT scan image dataset organized into multiple class-specific folders. The images are loaded using OpenCV and resized to a fixed resolution of 128×128 pixels to maintain a consistent input size. The images are subsequently converted from RGB format to grayscale, reducing the computational complexity while preserving important structural information required for classification.

Median Filtering is then applied to the grayscale images using a 5×5 kernel. This preprocessing step helps reduce noise and smooth the image while preserving important boundaries and structural details. The processed images provide a cleaner representation for subsequent feature extraction.

B. Frequency-Domain Feature Extraction

After preprocessing, Fast Fourier Transform (FFT) is applied to the CT scan images to transform the image representation from the spatial domain into the frequency domain. The FFT output is shifted using `fftshift` so that the low-frequency components are positioned at the center of the frequency spectrum.

A logarithmic magnitude representation is then generated from the transformed image. This frequency-domain representation captures information about the image's intensity variations and structural patterns. The resulting feature representation is flattened into a one-dimensional vector and used as input to the machine learning classifiers.

C. Machine Learning Classification

The proposed system evaluates multiple supervised machine learning algorithms for classification of the processed CT scan images. Four classifiers are incorporated into the framework:

- Support Vector Machine (SVM)
- Random Forest (RF)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN)

The extracted feature vectors and their corresponding class labels are divided into training and testing subsets using an 80:20 split with stratification. Each classifier is trained independently on the training data and evaluated on the testing data using Accuracy, Precision, Recall, F1 Score, and AUC-based analysis.

D. Prediction and Model Evaluation

Following model training, the trained classifiers are saved for subsequent prediction. For a new CT scan image, the same preprocessing and feature extraction operations are performed before the image is passed to the trained model. The classifier then predicts the corresponding lung cancer class.

The proposed framework also generates confusion matrices, ROC curves, classification reports, and comparative performance graphs to analyze the behavior of the different classifiers. Based on the experimental results obtained from the project implementation, the classification models can be compared using their measured performance metrics, providing a systematic basis for selecting a suitable model for lung cancer image classification.

V. MODULE DESCRIPTION

The proposed system consists of six interconnected modules that together form the complete medical image processing and lung cancer classification framework. Each module performs a specific operation, beginning with CT scan image acquisition and preprocessing and continuing through feature extraction, machine learning classification, evaluation, and final prediction.

A. Dataset and Image Loading Module

The dataset and image loading module is responsible for loading CT scan images from the class-organized dataset. The images are read using OpenCV and their corresponding class labels are assigned according to the directory structure. The module verifies the availability of valid image files and stores the images and labels for further processing. This module provides the initial input to the proposed lung cancer detection pipeline.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Image Preprocessing Module

The image preprocessing module prepares the input CT scan images for feature extraction and classification. Each image is resized to 128×128 pixels and converted into grayscale to maintain a uniform image representation and reduce computational complexity. Median Filtering with a 5×5 kernel is subsequently applied to reduce noise while preserving important structural information in the CT scan images.

C. Frequency-Domain Feature Extraction Module

The frequency-domain feature extraction module applies Fast Fourier Transform (FFT) to the preprocessed grayscale images. The transformed frequency spectrum is shifted using fftshift, and a logarithmic magnitude representation is generated to visualize and represent the frequency components of the image. The resulting feature representation is flattened into a one-dimensional feature vector, which is then provided as input to the classification models.

D. Machine Learning Classification Module

The machine learning classification module performs supervised classification of the extracted CT scan features. Four machine learning algorithms are implemented and evaluated: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN). The dataset is divided into training and testing subsets using an 80:20 stratified split. Each classifier is trained independently and used to classify the test samples into the corresponding lung cancer categories.

E. Performance Evaluation Module

The performance evaluation module analyzes and compares the results obtained from the different classification algorithms. The models are evaluated using Accuracy, Precision, Recall, F1 Score, and AUC-based analysis. Confusion matrices and ROC curves are also generated to examine classification performance across different classes. The results are stored in structured output files and graphical representations for comparative analysis.

F. Prediction and Model Deployment Module

The prediction module uses the trained and saved machine learning models to classify new CT scan images. A new image is first converted to grayscale, resized to 128×128 pixels, and processed using Median Filtering and FFT-based feature extraction. The resulting feature vector is passed to the selected trained classifier, which generates the predicted lung cancer class. The prediction result is then displayed to the user along with the processed input image.

VI. SIMULATION RESULTS AND ANALYSIS

The proposed lung cancer classification framework was experimentally evaluated using the CT scan image dataset and four machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN). The dataset contained 613 images belonging to four classes. The extracted feature vectors had a dimensionality of 16,384. An 80:20 stratified train-test split was used, resulting in 490 training samples and 123 testing samples.

TABLE I

EXPERIMENTAL PARAMETERS	
Parameter	Value
Dataset Images	613
Number of Classes	4
Image Size	128×128
Image Type	Grayscale
Median Filter Kernel	5×5
Feature Dimension	16,384
Training Samples	490
Testing Samples	123
Train-Test Split	80:20



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

SVM Kernel	Linear
Random Forest Estimators	100
KNN Neighbors	5
Random State	42

A. Dataset and Preprocessing Analysis

The dataset loading process successfully identified 613 CT scan images belonging to four different classes. Each image was resized to 128×128 pixels and converted into grayscale format. Median Filtering was subsequently applied to reduce noise and improve the quality of the input images.

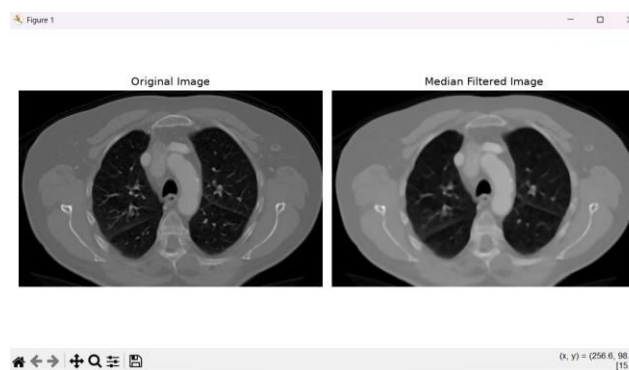


Fig. 1. Original and Median Filtered CT Scan Image

The preprocessing stage provides a standardized representation of the CT scan images before feature extraction. The application of Median Filtering produces a smoother image while retaining important structural information required for subsequent classification.

B. FFT Feature Extraction Analysis

Fast Fourier Transform was applied to the preprocessed grayscale images to obtain their frequency-domain representation. The FFT magnitude spectrum was generated after shifting the frequency components toward the center.

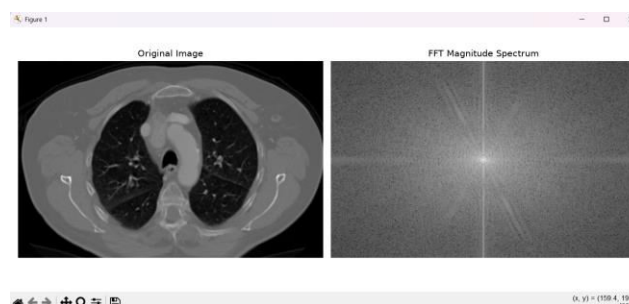


Fig. 2. Original CT Scan Image and FFT Magnitude Spectrum

The resulting frequency-domain representation was converted into a feature vector. The final feature representation contained 16,384 features for each image, which were subsequently used as input to the machine learning models.

C. Classification Accuracy Comparison

The four implemented machine learning algorithms were trained and evaluated using the same feature representation and testing dataset. The obtained classification accuracies are presented in Table II.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE II

CLASSIFICATION COMPARISON		PERFORMANCE
Algorithm	Accuracy	
SVM	77.24%	
Random Forest	69.92%	
Decision Tree	61.79%	
KNN	76.42%	

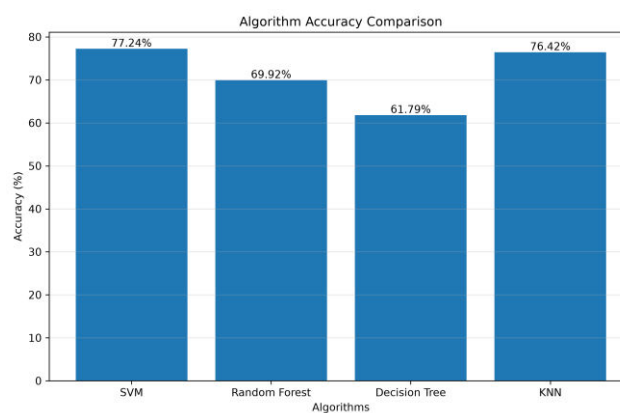


Fig. 3. Accuracy Comparison of Machine Learning Algorithms

Among the evaluated models, SVM achieved an accuracy of 77.24%, followed by KNN with 76.42%. Random Forest achieved 69.92%, while Decision Tree obtained 61.79%. The results indicate that the linear SVM provided the highest classification accuracy among the four evaluated models for the experimental configuration used in this study.

D. Precision, Recall and F1-Score Analysis

In addition to accuracy, Precision, Recall, and F1 Score were calculated to provide a more detailed evaluation of the classification models.

For SVM, the obtained Precision was 77.33%, Recall was 77.24%, and F1 Score was 75.70%. Random Forest achieved a Precision of 75.88%, Recall of 69.92%, and F1 Score of 66.64%.

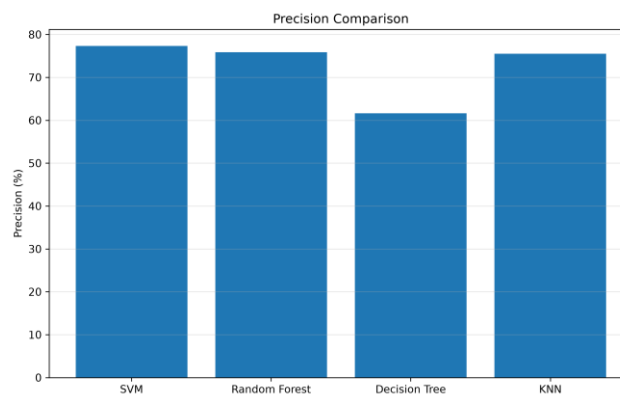


Fig. 4. Precision Comparison of Machine Learning Algorithms



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The results show that the SVM model maintained relatively balanced Precision, Recall, and F1 Score values, indicating consistent classification performance on the evaluated testing samples.

E. ROC and AUC Analysis

Receiver Operating Characteristic (ROC) analysis was performed to further evaluate the classification behavior of the trained models. The AUC value was calculated using the model probability estimates or decision scores.

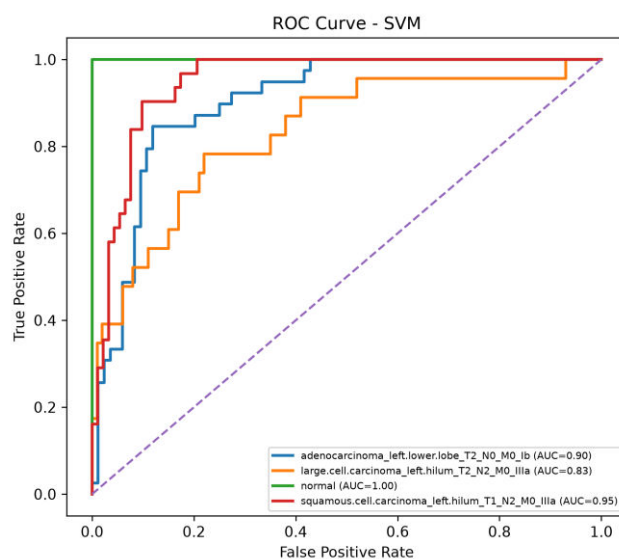


Fig. 5. ROC Curve Comparison of Classification Models

The ROC analysis provides an additional perspective on the ability of the classifiers to distinguish between the different lung cancer classes. The generated ROC curves and AUC values were stored as part of the experimental results for model evaluation.

F. Prediction Analysis

After training, the machine learning models were saved and used for prediction of new CT scan images. During prediction, the input image is converted to grayscale, resized to 128×128 pixels, processed using Median Filtering, and transformed using FFT before being passed to the trained classifier.

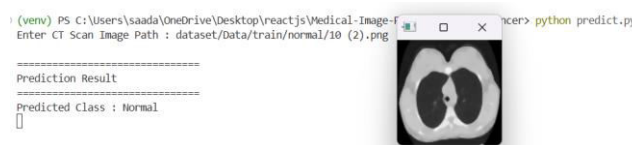


Fig. 6. Sample Lung Cancer Prediction Output

The prediction module successfully produced class labels for test CT scan images, including classes such as Adenocarcinoma and Large Cell Carcinoma. This demonstrates the complete workflow from image preprocessing and feature extraction to final classification.

G. Summary of Results

The experimental results demonstrate that the proposed image processing and machine learning framework provides a complete pipeline for lung cancer image classification. The major observations are:

- 613 CT scan images were processed.
- Four lung cancer-related classes were considered.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- The final feature representation contained 16,384 features per image.
- 490 images were used for training and 123 images for testing.
- SVM achieved the highest observed accuracy of 77.24%.
- KNN achieved an accuracy of 76.42%.
- Random Forest achieved an accuracy of 69.92%.
- Decision Tree achieved an accuracy of 61.79%.
- SVM achieved Precision of **77.33%**, Recall of **77.24%**, and F1 Score of **75.70%**.

These results demonstrate the effectiveness of combining image preprocessing, FFT-based feature extraction, and supervised machine learning for automated classification of lung CT scan images.

VII. CONCLUSION AND FUTURE WORK

The proposed medical image processing framework presents a machine learning-based approach for automated classification of lung cancer from CT scan images. The system integrates image preprocessing, Median Filtering, Fast Fourier Transform (FFT)-based feature extraction, and supervised machine learning algorithms to develop a complete classification pipeline. The experimental dataset consisted of 613 CT scan images belonging to four classes, with 490 images used for training and 123 images used for testing.

Four classification algorithms, namely Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN), were implemented and evaluated. Among the evaluated models, SVM achieved the highest classification accuracy of 77.24%, followed by KNN with 76.42%, Random Forest with 69.92%, and Decision Tree with 61.79%. The SVM model also achieved 77.33% Precision, 77.24% Recall, and 75.70% F1 Score. These results demonstrate that the combination of image preprocessing, frequency-domain feature extraction, and machine learning can provide a structured approach for lung cancer image classification.

The developed system can assist in the automated analysis of CT scan images and can serve as a foundation for further research in computer-aided lung cancer detection. However, the current framework is intended as a classification research prototype and should not be considered a replacement for clinical diagnosis.

Future work can focus on increasing the size and diversity of the dataset, applying advanced deep learning and transfer learning techniques, improving feature extraction methods, performing extensive hyperparameter optimization, and validating the system using independent clinical datasets. Further improvements in these areas may help enhance classification performance and improve the robustness and generalizability of the proposed framework.

REFERENCES

- [1] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2018.
- [2] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.
- [3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [4] T. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [5] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 2000.
- [6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.
- [7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [8] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Technical Report MSR-TR-98-14, 1998.
- [9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details