MACHINE LEARNING MODELS AND INNOVATIVE FEATURE DESIGN FOR THE CLASSIFICATION OF LUNG CANCER

Srimahalakshmi Balamurugan[1], Tamilarasan Pari[2], Thyagarajan K[3], P. Karthikeyan[4], Shivashiga A M[1], Selin Frajja S[1], Atthi Sushrutha A[1]

[1]Department of Artificial Intelligence and Machine Learning, Chennai Institute of Technology, Chennai-600069, India.
[2] Assistant Professor, Department of Mechatronics, Chennai Institute of Technology, Chennai-600069, India.
[3] Assistant Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai-600069, India.
[4] Associate Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai-600069, India.

***Abstract* - Lung cancer accounts for the majority of cancer-related deaths worldwide because of its high incidence and usually delayed diagnosis. It affects public health significantly and results in noteworthy rates of morbidity and mortality. A patients quality of life and chance of survival depend on prompt diagnosis and effective care. Currently computer vision and image processing methods are very beneficial for classifying lung cancer. By combining machine learning and manually created features the model presented in this paper effectively classifies lung cancer from CT scan images. The model begins by applying Gaussian filtering (GF) to preprocess and improve the quality of the input images. After that an image slice segmentation technique is used to accurately identify the diseased areas of the images. The Oriented Rapid and Rotated Abstraction (ORB) and Gray Level Co-occurrence Matrix (GLCM) methods are used to extract features. Finally using a Random Forest (RF) classifier the right classifiers for experimental lung cancer images are found. The efficacy and efficiency of the proposed method are evaluated against other current methods using a dataset of CT images of lung cancer. The recommended model showed excellent accuracy and efficiency scoring a 95. twenty-three percent. The experimental results which show the models superiority in a number of performance metrics illustrate its potential for practical application in lung cancer diagnosis and treatment planning.**

Keywords: Lung cancer, CT scan images, ORB feature extraction, GLCM feature extraction, Random Forest classification.

I. INTRODUCTION

Lung cancer is a type of cancer that occurs when cells in the lungs grow abnormally and without treatment. A severe ailment that poses heightened danger of seriousness. The symptoms of lung cancer include a persistent cough, chest pain, and trouble breathing. The IARC and ACS have released new information indicating that lung cancer will result in approximately8 million deaths in 2022, which is about 18 percent of all cancer deaths worldwide. There are several reasons why lung cancer is a significant cause of death. It is common for the disease to be diagnosed at an advanced stage, which limits the range of treatment options. Smoking is a major risk factor that accounts for approximately 85 percent of lung cancer cases. Excessive exposure to secondhand smoke, occupational hazards like asbestos and radon, air pollution trigger risk factors, and genetic factors. Globally, lung cancer is the most frequently diagnosed cancer, with almost 2.5 million new cases reported in 2022, representing approximately 12.4% is the proportion of cancer diagnoses (Homepage - IARC).

Preventing and controlling the progression of lung cancer requires early detection and classification. Traditional approaches to classifying the type of lung cancer have traditionally focused on assessing risk factors such as age, gender, and blood pressure and smoking habits. The emergence of machine learning and artificial intelligence has led to an increased interest in building lung cancer classification models that incorporate information from CT scan data.

In this manuscript, a new supervised model approach for early classification of lung cancer using images from CT scans is presented. A model usually has two main steps: feature extraction and classification. The extraction of features in the feature extraction step involves processing information from all shapes to identify features that can be categorized. For example, features such as morphological features,

intensity-based features, and geometric features can be extracted from CT scan data.

The input image is improved through the use of Gaussian filtering (GF) technique, which is used to preprocess the proposed model. In the lung image, a Grabcut-based segmentation technique is utilized to extract affected parts. Two other types of feature extractions are ORB feature extract and GLCM features extraction. Finally, a random forest (RF) model is used as the classification model.

II. RELATED WORKS

M. Abdar et al. [1] The paper introduced an approach that enhances the effectiveness of customary machine learning techniques employed in the study to anticipate CAD outcomes among patients in their Z-Alizadeh San dataset. This can lead to highly accurate models for both clinical and scientific use. P. Ratta et al. [2] New and innovative information technologies, such as IoT and Blockchain, have revolutionized health systems by improving their functionality, with three primary focus areas being drug traceability, remote patient monitoring, and disease management. B. From Potter et al. [3] The authors examined common thoracic, neurological, and musculoskeletal emergencies in individuals with lung cancer patients, including hypersensitivity syndromes. Mr. Vishal Patil et al. [4] Established a suitable image segmentation algorithm for medical images to restrict the ability of doctors in reading computed Tomography images. Suren Makaju et al. [5] Presented several computational methods, evaluated the existing best techniques and identified their limitations and inadequacies as well as a new model that enhances the current best model. Tang et al. [6] The conversation revolved around the topic of AlexNet, which seeks to reduce overfitting and improve uniformity while preventing gradient disappearance and explosion. Sethy et al. [7] Created a hybrid network for the classification of lung histopathology images using AlexNet, wavelet, and support vector machines. Pradhan. K et al. [8] Introduced diverse machine learning techniques utilized for detecting various illnesses to identify gaps and enhance lung cancer detection in the Medical Internet of Things. Yawei Li et al. [9] Provided an overview of machine learning-based techniques that can be applied to improve lung cancer diagnosis and treatment, with a focus on early detection, diagnosis, prognosis, and immunotherapy practice. Svoboda E [10] Suggested early prediction of lung cancer using computed tomography (CT) scan images through a supervised learning model system. Bhinder B et al. [11] Developed an algorithm for cancer identification and classification, molecular analysis of tumors with their microenvironment assay, drug discovery and redistribution, and patient prediction. Ocampo P et al. [12] stressed the importance of using convolutional neural networks (CNN) to analyze digital pathology slides and improve the accuracy of categorizing cancers, identifying specific genetic mutations that can improve NSCLC patients' diagnostic precision and personalised treatment. Ren Z et al. [13] The use of machine learning techniques has been augmented by deep learning methods to optimize the identification of lung cancer. By integrating multiple approaches, the study aims to use the strengths of each technique to improve the efficiency and reliability of lung cancer classification, which can contribute to better patient outcomes. Mathios D et al. [14] examined how patterns of cfDNA could be used as biomarkers for lung cancer to provide an early, non-invasive approach to detection and monitoring. Sunila Anjum et al. [15] By utilizing available resources, the aim was to expand the CNN model beyond just one element in depth and width while maintaining respect for resolution. Iftikhar Naseer et al. [16] Created highly effective techniques for block segmentation, candidate nodule removal, and lung cancer classification that enhanced accuracy. Tehnan I. A. Mohamed et al. [17] predicted the use of a new hybrid algorithm that could classify lung cancer more accurately, using – perhaps — the CNN model. Nusraat Nawreen et al. [18] Utilized thresholding and edge detection to segment the ROI area of lung tumor. Finally, they calculated several geometric features of the extracted ROIs and classified them into benign and malignant severities using a support vector machine (SVM) classifier. Sneha Balannolla et al. [19] Introduced a technique for detecting and categorizing lung nodules (or lesions) through utilizing multiple strategies.There are two aspects to it: nodule detection (finding of nodles) and classification (classifying nodes as benign/non-cancerous or malignant/cencerous). Imran Shafi [20] By using cross-sectional analysis, the proposed computer-aided design (CAD) model can be used to identify physiological and pathological changes in the soft tissues of lung cancer lesions.

III. THE PROPOSED MODEL

A detailed account of how the proposed model for categorizing lung cancer is presented, as demonstrated in Figure 1, Initially, the model uses a Gaussian filter technique to preprocess CT scan images. By

effectively filtering out noise and normalizing the image, this step significantly enhances the quality of the picture. This type of Gaussian filter smooths the image and makes subsequent steps in processing more accurate with a constant amount of noise. After that, Grabcut technology is used to segment the CT scan images. The potent segmentation technique distinguishes diseased and non-diseased areas in the lungs and isolates areas of significance for future investigation. Additionally, By accurately separating these components, the model can concentrate on providing accurate classification and diagnostics. Thus, the combination of noise-free Gaussian filtering and Grabcut's precise segmentation forms the backbone of the preprocessing step, ensuring that the images fed into the classification algorithm are of the highest quality and most meaningful.

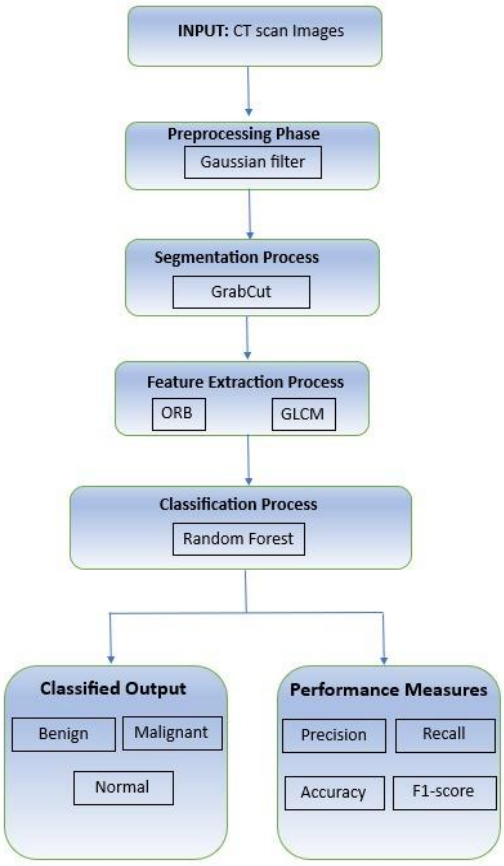


Fig. 1: Block diagram of Proposed model

Succeeded by Oriented FAST and rotated BRIEF (ORB) and Gray Level Co-Occurrence Matrix (GLCM) models are used for the extraction of meaningful features like texture, shape, intensity and statistical features which are essential for further examination.

3.1. Data Augmentation

In situations where it is challenging to obtain a large dataset, such as medical image analysis for lung cancer classification, data augmentation can be employed to increase the number of training samples and improve model precision. By artificially expanding the dataset, models can generalize better and avoid overfitting. Here we have optimized the geometric transformation such as rotation and flipping for augmentation process. It is either rotating the images by a certain range of degrees, such as ±20° or applying horizontal and vertical flips to increase the total number of images for obtaining large dataset. In this technique the new data samples are generated by applying various transformations to the existing dataset which reduce overfitting and enhance robustness. Further this large dataset is utilized to train the machine learning model. Increasing the amount of dataset improve the model performance by enhancing higher accuracy.

The material was expanded through a data addition process before the experimental validation. The procedure had an objective to add dimension to it. During this work, data is added in different ways:

- zoom_range=0.15,
- rotation_range=90,
- horizontal_flip=True,
- height_shift_range=0.2,
- fill_mode="nearest",
- shear_range=0.15,
- width_shift_range=0.2

3.2. Image Preprocessing using Gaussian Filter Technique

Image preprocessing plays a crucial role in lung cancer classification for early detection and accurate diagnosis by enhancing imaging and enabling automated screening. It aids in personalized treatment planning and monitoring by extracting vital features. The technique is used to preprocess the images and eliminate noise.GF is predominantly employed for image smoothing, noise removal, and image normalization to enhance its quality. The convolution operator is a Gaussian operator and the concept of Gaussian smoothing is implemented using convolutions.

The following is the definition of a 1-D Gaussian operator:

$$G_{1D}(x) = \frac{1}{\sqrt{2\pi}o} e^{-}\left(\frac{x^2}{2o^2}\right) \qquad (1)$$

The identification of the optimal smoothing filter for image processing necessitates a thorough analysis within both the spatial and frequency domains. This dual-domain approach ensures that the uncertainty principle is adequately addressed, thereby achieving the desired balance between image detail preservation and noise reduction, as given below:

$$\Delta x \Delta \omega \geq \frac{1}{2} \qquad (2)$$

The Gaussian operator in 2D can be represented by:

$$G_{2D}(x, y) = \frac{1}{2\pi\sigma^2} e^{-}\left(\frac{x^2 + y^2}{2\sigma^2}\right) \qquad (3)$$

where σ (Sigma) defines the standard deviation of the Gaussian function. Image smoothness is significantly enhanced with a higher value. The coordinates (x, y) of the image's rectangular coordinate points and the size of its filter window are also included. By using both additive and multiplicative operations between the kernel and the images, a matrix of values from 0 to 255 can be generated representing the image. A square normalized matrix with multiple bits makes up the kernel. The picture element is multiplied and the output is divided by two in the convolution process. The direct polynomial operation is used to accomplish this.

3.3. Image Segmentation using GrabCut Method

The preprocessed image is input and segmented for segmentation using the GrabCut process to see if there are any anomalies in the image. The recognized graph cut method a long-standing strategy in combinatorial graph theory has seen a notable increase in use. This method has been applied successfully by numerous researchers for image and video segmentation. The graph cutting method which applies a graph-cutting process is a more advanced technique for image segmentation. This method determines the background and foreground pixels by requiring user-defined character input. Making a graph that shows the level of correlation between background and foreground pixels is the process. Next the minimum crop needed to differentiate the foreground from the background is determined.

These are the characteristics of the energy function:

$$E(\underline{\alpha}, k, \underline{\theta}, z) = U(\underline{\alpha}, k, \underline{\theta}, z) + V(\underline{\alpha}, z) \qquad (4)$$

$$U(\underline{\alpha}, k, \underline{\theta}, z) = \sum_{n} D(\alpha_n, k_n, \underline{\theta}, z_n) \qquad (5)$$

$$V(\underline{\alpha}, z) = \gamma \sum_{(m,n)\varepsilon C} [\alpha_n \neq \alpha_m] \exp -\beta \|z_m - z_n\|^2 \qquad (6)$$

$$D(\alpha_n, k_n, \underline{\theta}, z_n) - \log p(z_n | \alpha_n, k_n, \underline{\theta}) - \log \pi(\alpha_n, k_n) \qquad (7)$$

In image segmentation the function U denotes the area-based component of the energy function. It calculates the probability that each pixel is in the foreground or background by utilizing a combination of Gaussian models. The accuracy of segmentation is improved by the Gaussian mixture model which quantifies the likelihood that a specific pixel is in the foreground or background.

The boundary-based component of the energy function on the other hand is represented by function V. To account for the discontinuities between neighboring pixels m and n it has a penalty term. A smaller variance suggests a higher likelihood that the pixels are part of the same segment either background or foreground. This penalty term is inversely proportional to the variance between two neighboring pixels. On the other hand a higher variance implies that the pixels are more likely to belong to distinct segments and be on opposing sides of an edge.

The method effectively computes the probability of each pixel being in the foreground or background by utilizing a Gaussian mixture model. By integrating both the area-based and boundary-based components the overall energy function is optimized to achieve the final segmentation. Through this optimization process it is made sure that the image is segmented in a way that preserves meaningful and cogent boundaries between various segments while optimizing the likelihoods obtained from the Gaussian models.

3.4. Feature Extraction

3.4.1 ORB based Feature Extraction

Using features from both LONG descriptors and FAST keypoint detectors the ORB technique is used in the feature extraction process. This method includes a number of changes intended to improve performance. The accuracy and speed of feature extraction are greatly increased by the ORF technique which combines the strength and compactness of

BRIEF descriptors with the efficiency of FAST keypoint detectors which are excellent at rapidly identifying important points in images. These enhancements ensure that the ORF technique captures more relevant and distinct features, thereby increasing the robustness and effectiveness of feature extraction in various applications such as image recognition and object detection.

$$m_{pq} = \sum_{x,y} x^p y^q I(x,y) \qquad (8)$$

Where m_{pq} refers to the $(p+q)^{th}$ order moment of an image and the intensity $I(x,y)$ differs as a function of x and y image coordinate. By considering the moments in Eq. (6), the centroid is attained by:

$$C\left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \qquad (9)$$

A vector is generated from the centre to centroid $\overrightarrow{OC}$ and later the orientation of patch turns into as:

$$\theta = atan2(m_{01}, m_{10}) \qquad (10)$$

where $atan2$ represents the quadrant aware version of arctan. The influence of the illumination variable at corners is considered negligible, as the angle measurement remains constant regardless of the corner's type or configuration. To enhance rotational invariance, it is essential to calculate the moment with respect to the X and Y coordinates within a circular region of radius r. Selecting an optimal patch size of r ensures that the x and y coordinates range from [-r, r]. Typically, when using the Hessian measure, the value of |c| becomes 0, leading to instability. However, this issue does not arise with the FAST algorithm, which significantly benefits system efficiency. The stability and efficiency improvements provided by FAST make it a preferable choice over the Hessian measure for achieving rotational invariance in various applications. The ORB (Oriented FAST and Rotated BRIEF) algorithm incorporates a rotation-aware module known as r-BRIEF, which is an advanced version of the steered BRIEF descriptor. This improvement is combined with an appropriate learning phase intended to find binary features that are less correlated. An image patchs bit-string representation is created using a series of binary intensity tests to guarantee the BRIEF function rotates effectively. ORB is a reliable option for a range of computer vision applications because this method improves rotational invariance and feature matching accuracy.

With the assumption that a smooth image patch p exists an orientation module is integrated into ORB to improve the conventional BRIEF procedure. A binary test is also provided as.

$$\tau(p;x,y) := \begin{cases} 1: & p(x) < p(y) \\ 0: & p(x) \geq p(y) \end{cases} \qquad (11)$$

After considering the patch which is then represented as a vector of n binary tests as described below the feature is processed. Here p(x) denotes the intensity of the patch at a given point x.

$$f_n(p) := \sum_{1 \leq i \leq n} 2^{i-1} \tau(p; x_i, y_i) \qquad (12)$$

In order to get moderate results a Gaussian distribution is used in the vicinity of the patchs center and the vector length n is set at 256. The initial phase of this method is called steered BRIEF and it entails guiding BRIEF using keypoint orientation. The way this works is that the orientation of the keypoints is used to steer the BRIEF descriptor for a given feature set that contains n binary tests at particular positions. This ensures better accuracy and robustness in feature representation. A matrix representing the binary test at a particular position is provided by:

$$S = \begin{pmatrix} x_1, & \cdots & x_n \\ y_1, & \cdots & y_n \end{pmatrix} \qquad (13)$$

Subsequently, by employing θ (patch orientation) and R_θ (equivalent rotation matrix), a steered version $S_\theta$of S is reached as follows:

$$S_\theta = R_\theta S \qquad (14)$$

Therefore, the steered BRIEF function is given by:

$$g_n(p, \theta) := f_n(P) | (x_i y_i) \in S_\theta \qquad (15)$$

The angle is discretized in (twelve degree) increments and a lookup table of previously calculated BRIEF patterns is produced. It is possible to accurately compute the BRIEF descriptor by using the precise set of points as long as the keypoint orientation is maintained across various perspectives. By using this method the feature representation becomes more robust and reliable guaranteeing that the descriptor will hold true and remain accurate even when subjected to different transformations and points of view.

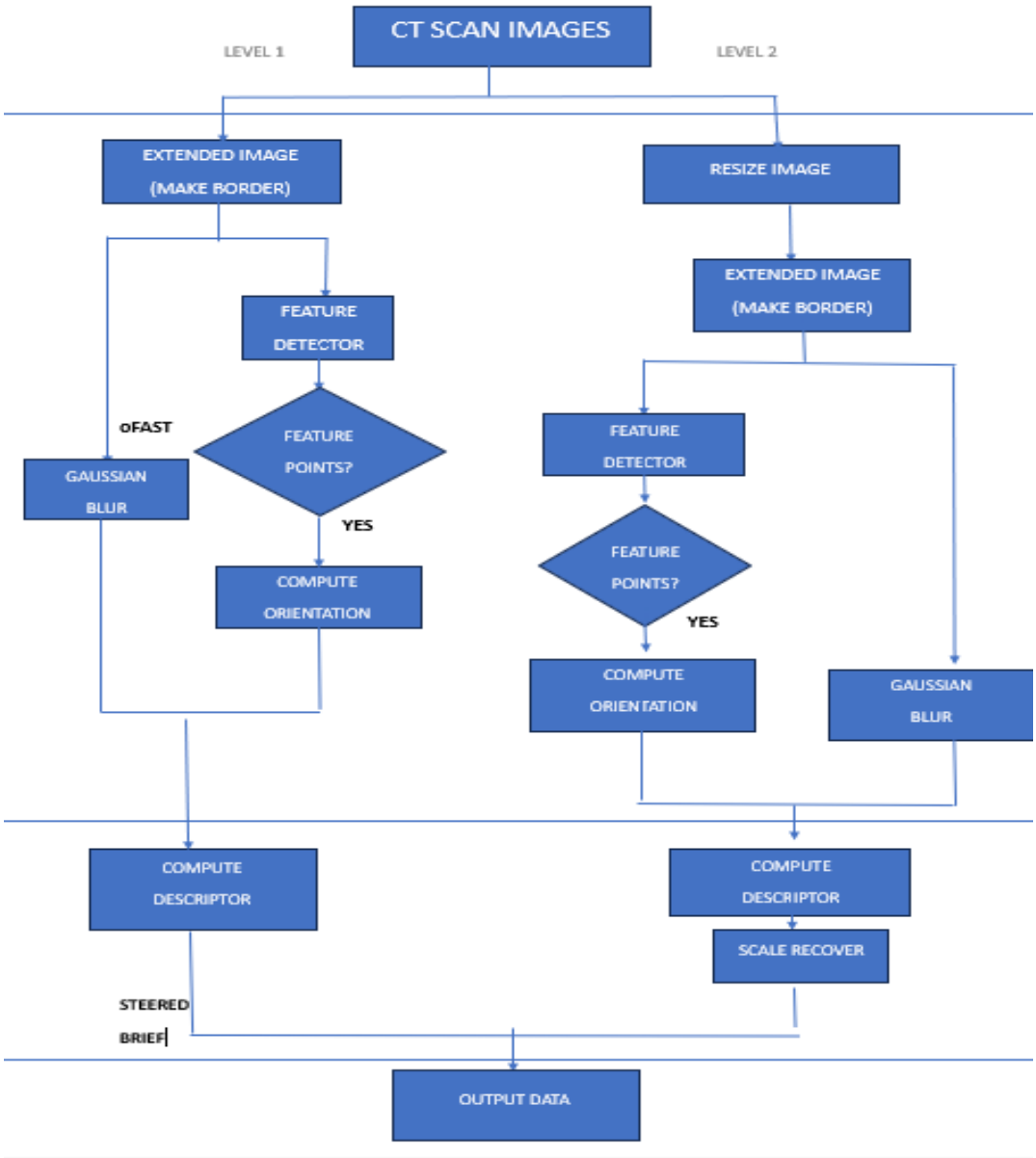


Fig.2. Process of ORB feature extractor

3.4.2 GLCM based Feature Extraction

The statistical technique known as the Gray Level Co-occurrence Matrix or GLCM examines textures by taking into account the spatial relationships between individual pixels. By computing pixel pairs with particular parameters and within a given spatial variation in the image GLCM functions characterize the texture of an image. Statistical measures are extracted from this matrix and a GLCM feature is created in order to accomplish this. By looking at variations of gray levels V and U at a desired distance D and a defined angle . the GLCM determines the spatial relationship of pixels at each intensity. This method enables a thorough understanding of the texture based on the spatial distribution and intensity of pixel pairs.

Contrast: It acts as a gauge for the overall images intensity contrast between a pixel and its surrounding pixels. The coordinates of the gray level pixels are represented by the co-occurrence matrix which looks like this:

$$Contrast = \sum_{U,V} |U - V|^2 P(U,V) \qquad (16)$$

Here an element in the co-occurrence matrix is denoted by P (U V) where U and V stand for the gray level pixels.

Correlation: It functions as a gauge of how closely adjacent pixels correlate with one another throughout the whole picture.

$$Correlation = \sum_{U,V} \frac{(\mu - \mu_U)(V - \mu_V)P(U,V)}{\sigma_U \sigma_V} \quad (17)$$

Here, $\sigma_U \sigma_V$ is referred as a standard deviation of $U$and V, μ_V is defined as the mean of U, μ_U is represented as the mean of V.

Energy: It acts as a data measure. Energy can be used as a positive gauge that should be maximized or as a negative gauge that should be minimized. The GLCM quantifies the total squares of the associated variables. The formula below is used to calculate the energy:

$$Energy = \sum_{U,V} P(U,V)^2 \quad (18)$$

Homogeneity: This formula is used to calculate the degree to which the variable distribution in the GLCM is near the diagonal.

$$Homogeneity = \sum_{U,V} \frac{P(U,V)}{1 + |I - J|} \quad (19)$$

3.5. Image Classification

Random Forest Classifier

For classification and regression issues the Random Forest model is used as an alternative to ensemble learning. During training it produces a number of decision trees from which it extracts a class mode for classification or an average prediction for regression analysis. This method considers a random subset of features at each split and combines bagging (bootstrap aggregating) with random feature selection. Each tree is trained on a different subset of data. By averaging out biases and decreasing variance this procedure helps to improve generalization and lessen overfitting. Large datasets with high dimensionality can be handled by the scalable reliable Random Forest model. Furthermore the model provides insights into feature importance that are helpful in comprehending the data and the decision-making process of the model. Because of their general strength and versatility Random Forests are a popular option for many machine learning applications.

Process:

- **Bootstrapping:** Random Forest generates multiple subsets of the training data using bootstrapping techniques.
- **Tree Construction:** For each one of the subsets, a decision tree is constructed. During this process:
  - Random Forest chooses a random subset of features at each split point.
  - Entropy and other factors like Gini impurity are used to determine which split is the best.
- **Aggregation:** The final prediction is made by combining the predictions of all the trees. Given T decision trees, the Random Forest prediction for an input x is:

$$\hat{y} = majority_{vote\{h_t(x)\}_{t=1}^{T}} \quad (20)$$

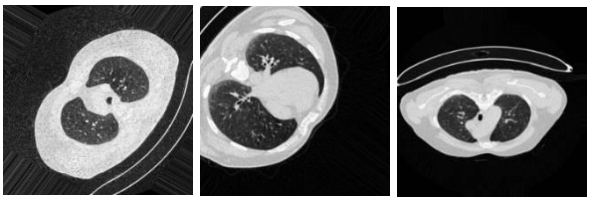
where $h_t(x)$ is the prediction of the t[th] decision tree.

IV. PERFORMANCE VALIDATION

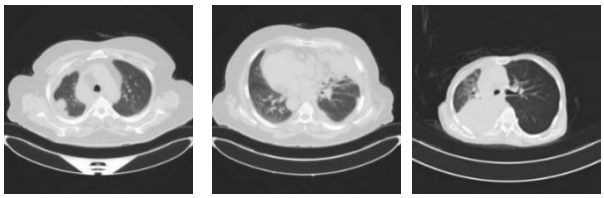
The proposed model's performance is tested using the benchmark dataset. The method is simulated using Python 3.6.5. Here the Table 1 provides dataset details, and Fig. 3 shows sample test images. The dataset holds a set of 221 images under Benign case, 561 images under Malignant case and 416 images under Normal case. The sample processes that are obtained during simulation is provided in Appendix.

Table 1 – Dataset Description

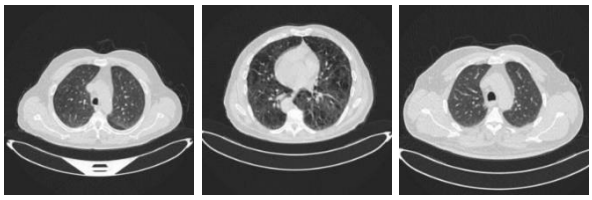
Classes	Number of images
Benign	221
Malignant	561
Normal	416
Total	1198



(a)



(b)



(c)

Fig. 3(a) Benign case, **3(b)** Malignant case, **3(c)** Normal case

In the comparative analysis of the GLCM-RF (Gray Level Co-occurrence Matrix with Random Forest) model versus the ORB-RF (Oriented FAST and Rotated BRIEF with Random Forest) model, the GLCM-RF model generally outperforms due to its superior ability to capture texture features in CT scan images, which are crucial for distinguishing between different stages of lung cancer. GLCM provides detailed statistical measures of image texture, such as contrast, correlation, energy, and homogeneity, which are particularly effective in identifying subtle differences in the patterns and structures of lung tissues. On the other hand, ORB focuses on detecting key points and their descriptors, which, while effective for tasks like object recognition and image matching, may not be as robust in capturing the intricate textural variations present in medical images. Consequently, the GLCM-RF model tends to achieve higher accuracy and reliability in classifying lung cancer stages, as it leverages more comprehensive and relevant feature extraction tailored to the specific characteristics of CT scan images. Apparently, the ORB-RF model has accomplished the least performance of 55.83% whereas the GLCM-RF models has accomplished the higher performance of 95.23%. Therefore, among the two proposed models, the GLCM-RF model is found to be superior and appeared as an effective lung cancer classification model.

Table 3 – Comparative analysis of the two proposed model

Model	Accuracy
ORB-RF	55.83
GLCM-RF	95.23

The confusion matrix produced on the classification of lung cancer by ORB-RF model is given in Fig. 4(a) and that of GLCM-RF model is given in Fig. 4(b).

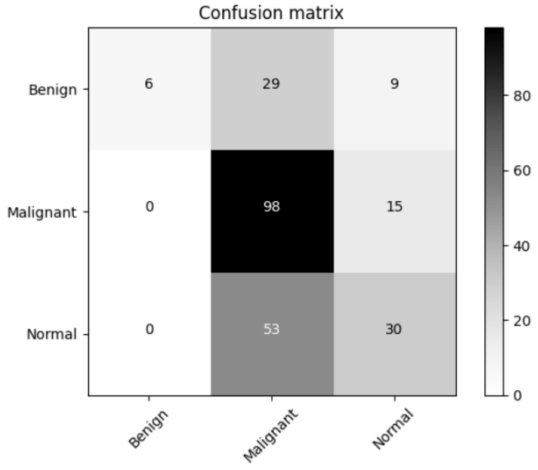


Fig. 4(a) Confusion Matrix of ORB-RF model

From the figure 4(a) it can be observed that the ORB-RF approach has classified a set of 29 images as Benign, 98 images under Malignant and 53 images into Normal cases. This model accomplishes an accuracy rate of 55.83 percentage.

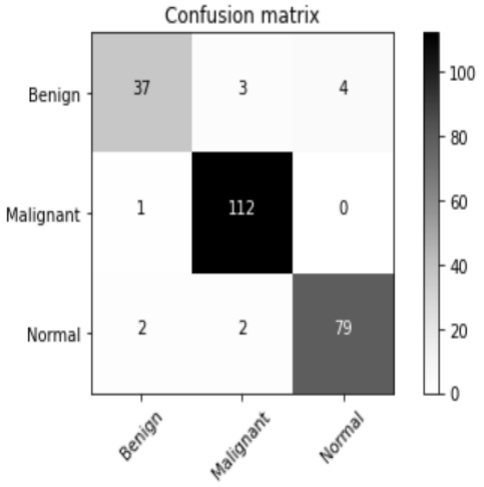


Fig. 4(b) Confusion Matrix of GLCM-RF model

In Figure 4 (b), the GLCM-RF technique identified 37 image sets as benign, 112 images as malignant, and 79 images were classified as normal cases using different methods. This model is 95.23 percentage times more accurate than others.

Table 3 and Figure 5 display the proposed models' performance in classifying lung cancer using test images. The proposed random forest classification model classified lung images with an accuracy of 95.23, precision of 94.67, recall score of 92.82, and F1 score of 93.34.

Table 3 - Result Analysis of Proposed Methods in terms of Different Measures

Methods	Accy.	Precision	Recall	F1 - Score
SVM	88.23	89.33	87.49	88.59
Decision Tree	90.31	91.45	88.78	90.82
Logistic Regression	93.19	92.87	90.17	91.76
RF Model	95.23	94.67	92.82	93.34

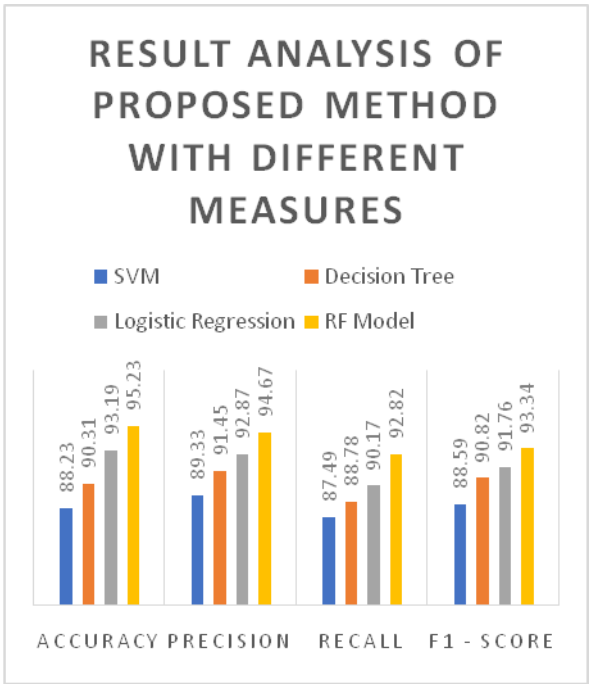


Fig. 5. Result analysis of proposed method with different measures

The represented Decision Tree model has effectively classified the lung cancer disease images with an accuracy of 90.31, precision of 91.45, recall of 88.78 and F1 – Score of 90.82. Furthermore, the utilized Logistic Regression model classified the images with an accuracy of 93.19, precision of 92.87, recall of 90.17 and F1 – Score of 91.76. Moreover, the presented Random Forest classifier methodology classified the lung cancer disease images with an accuracy of 95.23, precision of 94.67, recall of 92.82 and F1 – Score of 93.34.

Algorithm 1 Lung Cancer Classification Using Image Processing and Machine Learning

Input: Dataset of CT scan images
Output: ORB-RF Model Accuracy, GLCM-RF Model Accuracy
Load dataset
for each image **in** dataset **do**
    image ← **Apply** Gaussian Filter **and** Augment image
end for
for each image **in** dataset **do**
    Segmented_image ← **Apply** GrabCut Technique **to** image
end for
features_ORB ← **Extract** ORB Features **from** segmented_images
features_GLCM ← **Extract** GLCM Features **from** segmented_images
model_ORB_RF ← **Train** Random Forest Classifier **using** features_ORB
accuracy_ORB_RF ← **Evaluate** model_ORB_RF
model_GLCM_RF ← **Train** Random Forest Classifier **using** features_GLCM
accuracy_GLCM_RF ← **Evaluate** model_GLCM_RF
Print "ORB-RF Model Accuracy: " accuracy_ORB_RF " %"
Print "GLCM-RF Model Accuracy: " accuracy_GLCM_RF " %"

V. CONCLUSION

This paper introduces an advanced hand-crafted feature and machine learning-based model for lung cancer classification. The proposed methodology leverages the Gaussian Filter technique for image preprocessing, effectively reducing noise. Subsequently, the Grabcut algorithm is employed to segment diseased and non-diseased regions within lung images. To extract meaningful features, the model utilizes ORB (Oriented FAST and Rotated BRIEF) and GLCM (Gray Level Co-occurrence Matrix) techniques, which are crucial for further analysis. Ultimately, the Random Forest Classifier is applied to categorize lung images into three classes: Benign, Malignant, and Normal. To evaluate the classification performance of the proposed model, a series of simulations were conducted on a benchmark lung cancer dataset. The experimental results indicate that the presented method outperforms recent techniques across various metrics. Looking ahead, the efficiency of the proposed approach could be further enhanced through the integration of deep learning models.

VI. REFERENCE

1. M. Abdar, W. Ksiaúzek, U.R. Acharya, R.-S. Tan, V. Makarenkov, P. Pawiak, “A new machine learning technique for an accurate diagnosis of coronary artery disease”, Computer methods and programs in biomedicine., 179 (2019).

2. P. Ratta, A. Kaur, S. Sharma, M. Shabaz, G. Dhiman, “Application of blockchain and internet of things in healthcare and medical sector: applications, challenges, and future perspectives”, Journal of Food Quality., 1 (2021).

3. B. De Potter, J. Huyskens, B. Hiddinga, “Imaging of Urgencies and Emergencies in the Lung Cancer Patient”, Springer, New York, NY, USA (2018).

4. Mr. Vishal Patil, Dr. Aditya Gupta, Mr. Avinash Pawar, “Lung Cancer Detection Using Image Processing”, Journal of Emerging Technology and Innovative Research., Volume 8, Issue 3 (2021).

5. Suren Makaju, P.W.C. Prasad, Abeer Alsadoona, A. K. Singh, A. Elchouemi, “Lung Cancer Detection using CT Scan Images”, Procedia Computer Science 125., 107–114 (2018).

6. Tang, W.; Sun, J.; Wang, S.; Zhang, Y, “Review of AlexNet for Medical Image Classification”, arXiv preprint arXiv:2311.08655, (2023).

7. Sethy, P.K.; Geetha Devi, A.; Padhan, B.; Behera, S.K.; Sreedhar, S.; Das, K, “Lung Cancer Histopathological Image Classification Using Wavelets and AlexNet”, Journal of X-Ray Science and Technology., 31, 211–221 (2023).

8. Pradhan, K.; Chawla, P, “Medical Internet of Things Using Machine Learning Algorithms for Lung Cancer Detection”, Journal of Management Analytics., 7, 591–623 (2020).

9. Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang and Yuan Luo, “Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis”, Genomics Proteomics Bioinformatics., 20(5): 850–866 (2022).

10. Svoboda E, “Artificial intelligence is improving the detection of lung cancer”, Nature., 587: S20–S22 (2020).

11. Bhinder B, Gilvary C, Madhukar N.S, Elemento O, “Artificial intelligence in cancer research and precision medicine”, Cancer Discov., 11:900–915 (2021).

12. Ocampo P, Moreira A, Coudray N, Sakellaropoulos T, Narula N, Snuderl M., et al. “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning”, Journal of Thorac Oncol., 13: S562 (2018).

13. Ren Z, Zhang Y, Wang S. A hybrid framework for lung cancer classification. Electronics., 11:1614 (2022).

14. Mathios D, Johansen J.S, Cristiano S, Medina J.E, Phallen J, Larsen K.R, et al. “Detection and characterization of lung cancer using cell-free DNA fragmentizes”, Nature communications., 12:5060 (2021).

15. Sunila Anjum, Imran Ahmed, Muhammad Asif, Hanan Aljuaid, Fahad Alturise, Yazeed Yasin Ghadi, Rashad Elhabob, “Lung Cancer Classification in Histopathology Images Using Multiresolution Efficient Nets”, Computational Intelligence and Neuroscience., (2023).

16. Iftikhar Naseer, Sheeraz Akram, Tehreem Masood, Muhammad Rashid and Arfan Jaffar, “Lung Cancer Classification Using Modified U-Net based Lobe Segmentation and Nodule Detection”, IEEE Access., (2023).

17. Tehnan I. A. Mohamed, Olaide N. Oyelade, Absalom E. Ezugwu, “Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm”, PLoS One., 18-8 (2023).

18. Nusraat Nawreen, Umma Hany, Tahmina Islam, et al. “Lung Cancer Detection and Classification using CT Scan Image Processing”, IEEE Access., (2021).

19. Sneha Balannolla, Dr. A. Kousar Nikhath, Dr. Sagar Yeruva et al. “Detection and Classification of Lung Carcinoma using CT scans”, Journal of Physics: Conference Series., 2286 (2022).

20. Imran Shafi, Sadia Din, Asim Khan et al. “An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network”, Cancers (Basel)., 14(21): 5457(2022).