International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 05 | May 2025 www.irjet.net p-ISSN: 2395-0072

DETECTION AND ANALYSIS OF LUNG CANCER TUMORS USING ML TECHNIQUES

P.S.Mayura Veena[1], M.Chandrika[2], Akshaya.T[3], S.Karthik[4], Ch.Harshavardhan[5]

[1]*Assistant Professor, Department of ECE, Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India*
[2]*UG student, Department of ECE, Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India*
[3]*UG student, Department of ECE, Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India*
[4]*UG student, Department of ECE, Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India*
[5]*UG student, Department of ECE, Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh, India*

Abstract- *Lung cancer is currently the most frequently diagnosed major cancer and the most common cause of cancer mortality worldwide. Early and accurate detection results in more effective treatment options, thereby improving the patient outcomes. This study presents an automated classification model for lung cancer using machine learning techniques applied to CT scan images. At first, the dataset is divided into training , validation, and testing sets. The process begins with image pre-processing, followed by feature extraction using Histogram of Oriented Gradients (HOG) and classification using machine learning classifiers- Support Vector Machine (SVM), logistic regression, and Naive Bayes. Finally, the models are evaluated using various metrics like accuracy, precision, recall, and F1-score. Among the models, SVM achieved the highest training accuracy of 96.41%. Based on the classification results, only the images that are predicted as cancerous are passed on for segmentation, thereby reducing the computational load. The proposed model demonstrates its strong ability to assist in early detection and analysis of lung cancer tumors.*

Key words- Machine learning (ML), Histogram of Oriented Gradients (HOG), Naive Bayes, Support Vector Machine (SVM), Logistic Regression.

1.INTRODUCTION

Cancer is one of the main causes of non-accidental deaths, with lung cancer being its main contributor. Early detection enables timely and effective treatment, significantly improving the patient outcomes. The uncontrollable growth of abnormal cells forming malignant tumors leads to lung cancer. Lung cancer also has the ability to spread to the adjacent tissues and organs.

Lung cancer is broadly classified into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). This study focuses exclusively on NSCLC, which accounts for nearly 85% of all lung cancer cases. NSCLC is categorized into three major subtypes: adenocarcinoma, majorly found in the outer regions of the lungs, which arises from epithelial cells; squamous cell carcinoma, commonly located near the central bronchi; and large cell carcinoma, which typically occurs in any lung region and is characterized by rapid growth and metastasis.

Traditional diagnostic methods like analysis of lung CT scans by radiologists require expertise and may be limited by variability and subjectivity in interpretation. Early-stage nodules, i.e., subtle nodules, may be misclassified or overlooked, thereby leading to suboptimal treatment. Therefore, to avoid these limitations, an automated diagnostic tool that supports early and accurate detection of lung cancer is developed.

Machine learning provides a powerful alternative tool enabling automated, objective analysis of complex imaging data. In this study, three ML algorithms, like support vector machine (SVM), Logistic Regression, and Naive Bayes, are used for the multiclass classification of NSCLC subtypes using CT scan images.In examining the structural and textual features of lung tissues, the HOG method is employed.

Every algorithm offers specific benefits: SVM is suited for both linearly separable data and high dimensional spaces; logistic regression is easy to understand and offers a clear rationale; and Naive Bayes is commonly used because it is simple and computationally efficient. Using three ML models, their accuracy, precision, recall, and F1-score are used to evaluate and compare the algorithms. In the proposed diagnostic model, classification is used for initial lung cancer screening. If a CT image is classified as one of the subtypes of lung cancer by the model, it goes next to the segmentation of cancerous tumor nodule nodules. This stepwise strategy lessens computational demand, avoids superfluous segmentation, and maximizes resources, improving overall effectiveness and efficiency.

2.LITERATURE SURVEY

As noted earlier, lung cancer is one of the most common causes of cancer death all over the world which requires timely detection and classification. There is often a

misinterpretation with the conventional methods used for diagnosing such as analysis of CT Scan, giving rise to a need for automated diagnostic techniques.

Detection of lung cancer using Machine Learning algorithms has widely been practiced. A model was proposed by Hazra et al. [1] which applied support vector machine and logistic regression on a clinical data set of 422 lung CT scans. The research focused on lung cancer predictive modeling using advanced pre-processing methods such as multiple imputation and normalization, and subsequently, feature selection through Pearson correlation. Out of the two algorithms employed, logistic regression gave the best result with an accuracy value of 77.4% and SVM was able to achive 76.2% accuracy.

In the work proposed by Agarwal et al. [2], Random Forest was the best-performing algorithm among the studied machine learning methods for lung cancer detection and exposed an accuracy level of 92.3% followed by Decision Trees with 91% accuracy, SVM at 89.7% and Logistic Regression with 88.5% showing the importance of machine learning algorithms for predictive tasks.

Zhang et al. [3] employed support vector machines in detecting a subclass of lung cancer disorder, pneumoconiosis. This work revolves around improving models with custom tailored feature subsets.

Verma et al. [4] analyzed the application of several classification methodologies over different types of cancer. The analysis indicated that simple regression for quantile lung cancer detection attained an accuracy rate of 96.7 percent lung cancer quantile detection. This illustrates how machine learning algorithms perform well given a proper consolidated data arrangement.

Aljuaid et al. [5] applied Convolutional Neural Networks (CNN) for feature extraction, while classification was performed using Artificial Neural Network(ANN) and Naive Bayes algorithm. The ANN classifier achieved an accuracy of 95%, outperforming Naive Bayes, which achieved only 78%.

Lim et al. [6] and Avci [7] proposed a hybrid model that utilized SVM as a classifier, which is effective in detecting interstitial lung diseases.

Van Belle et al. [8] employed an SVM training strategy including support vector selection and polynomial kernels for classification of CT scans. This method is very useful when the data set contains high-dimensional medical images.

Weber et al. [9] employed FDG-PET/CT scan imaging to obtain chemotherapeutic responses while conducting a multicenter trial. This proposed work ensures a match between integrated imaging biomarkers and ML algorithm-based predictions.

These studies conclude that the ML algorithms make a huge mark in diagnosing lung cancer. However, the study still shows that there are many challenges to be addressed , such as data variability and subtype differentiation. Therefore, the model is built ensuring to avoid the above limitations and focus on precise and accurate detection of the lung cancer.

3.OUR PROPOSED WORK

3.1 Input Data

The input data includes the lung CT scan images , which have been collected from the local hospitals. All these CT scan images have been annotated by expert radiologists after effective analyses, ensuring high clinical accuracy. Each image is labeled as one of the four categories, including adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal lung.

The data set was divided into three subsets to ensure effective model development.

Training set: This part of the data set is used for training the model, enabling the model to learn discriminative features associated with diverse lung cancer conditions.

Validation set: This part of the data set is used for fine-tuning the hyperparameters, avoiding the risk of overfitting.

Test set: This part of the data set helps in developing the models generalization ability on unseen data, thereby ensuring unbiased assessment.

This form of data division ensures effective training and validation and testing, reflecting the real-world scenarios.

3.2 Pre-processing

The CT scan images were pre-processed to standardize and tune them for analysis prior to modeling.

All CT images were loaded onto the system with the OpenCV library. As these are filled with all the information from the scan, they will have to be processed later on grayscale images for simplicity and, most importantly, on major relevant structural features.

The principles of pre-processing are two:

Resizing*:* Resizing all images into a fixed size of resolution 128 by 64 pixels. The resemblance further on proved less in the run time of the processor and gave uniform images as input to the model.

Grayscale Conversion*:* The color images converted into grayscale to exhibit a texture and edge distinction as intensity variations. As it is viewed, the color information will not serve as useful while analyzing lung tissues in the CT, thus performing grayscale conversion allows one to focus mainly on significant density patterns and the structure.

Such pre-processing steps will improve and facilitate the extraction of features and simplify the entire process of learning in the model.

3.3 Feature Extraction

Feature extraction converts the raw input data into a set of measurable, indicative values, or features, that represent the most important characteristics of that data for a given task, such as classification, detection, or recognition. One of the most powerful techniques in this context is the Histogram of Oriented Gradients (HOG).

Histogram of Oriented Gradients (HOG): A feature descriptor widely employed in computer vision. HOG captures object shapes and structural information by doing analysis of edge directions in the localized regions in an image.

The steps include: First, the image will be pre-processed, which also consists of resizing to a standard size such as 128×64 pixels, and applying grayscale. Then, calculations of gradients are done to find pixels that would change significantly in intensity, resulting into edges. Then, the image is further divided into small cell sizes (from 8 by 8 or 16 by 16 pixel sizes) for every cell producing an orientation histogram based on the gradient direction weighted by its magnitude. This process includes several steps involving group cells in overlapping blocks to normalize the histograms made invariant to variations in lighting or contrast.The normative histograms are merged into one vector where the numbers describe different structural features of the image. Clearly, the CT scan data has been distilled into a format that can be processed using conventional machine learning algorithms like Support Vector Machine , Logistic Regression , or Gaussian Naïve Bayes , due to its numerical structure. The classifiers that were trained and evaluated with these features aim to distinguish adenocarcinomas from large cell carcinomas, squamous cell carcinomas, and lung conditions deemed normal.

3.4 Machine Learning Model Implementation

Support Vector Machine (SVM), Logistic Regression, and Naive Bayes, all widely known, were used for the classification problem in this research analysis. As a result of the well-defined features obtained through methods like HOG, the models were used to assess the performance of lung cancer image classification into adenocarcinoma, large cell carcinoma, normal, and squamous cell carcinoma.

Support Vector Machine (SVM): Among various supervised learning algorithms, the one that is most powerful is the Support Vector Machine (SVM) algorithm. It constructs a hyperplane that separates the classes of data in a high dimension feature space. The margin maximizes space between classes, thus enabling even very difficult datasets to be robustly classified.In comparison with the above discussed models, this study implemented a linear SVM model, which is effective when the feature space in terms of HOG features is linearly separable. The SVM classifier being trained with feature vectors obtained from HOG evaluated its performance using validation and test datasets on various evaluation metrics such as accuracy, precision, recall, F1-score and confusion matrix. These metrics provided full-fledged detail of how well the model was classifying the images into the respective categories. Therefore, with respect to its advantages in handling high-dimensional feature spaces and providing robust decision boundaries, SVM was expected to prove a convenient method for lung cancer classification.

Logistic Regression: Definition of logistic regression is that it is a probabilistic model that defines the terms in which the input features are mapped to the probability of an outcome falling into a class. The output is a probability from a logistic function, which is mapped to one of the classes. This time around, Logistic Regression was trained on the same HOG feature vectors and judged with the same metrics as SVM: accuracy, precision, recall, the F1 score, and confusion matrix. Unlike SVM, which focuses on finding an optimal separating hyperplane, this is probabilistic framework predicting not just class labels but also confidence values for each prediction regarding the output. This probabilistic output can really come in handy in the medical domain when one needs to understand how confident the system is about a certain diagnosis.

Naive Bayes: This classifier is based on Bayes' Theorem and assumes the conditional independence of the features given the class label. This assumption eases the classification task and makes Naive Bayes a computationally efficient algorithm. In this study, the Gaussian Naive Bayes classifier was applied, which assumes a Gaussian (normal) distribution of the features. This assumption fits very well with the continuous nature of the HOG features. After training the Naive Bayes classifier on the features extracted, validation, and testing of the classifier was performed. Evaluation of the performance was done using the same evaluation measures as for the other classifiers: accuracy, precision, recall, F1-score, and confusion matrix. Though Naive Bayes assumes conditional independence among the features, it remained competitive computationally, with ease of implementation arguing well in its favor.

3.5 Evaluation metrics

To evaluate the performance of the classification models, various metrics which are standardized such as precision, accuracy, recall, confusion matrix and F1-score are utilized. They help in providing us a deeper understanding about the usefulness and classification ability of lung cancer using HOG features through the models.

1. Accuracy: It says how accurate the model has made withthe following equation:

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP}$$

2. Precision: The overall depicted rate of positives is given by precision and the equation is:

$$Precision = \frac{TP}{FP + TP}$$

3. Recall: It gives the correctly predicted class with the following equation:

$$Recall = \frac{TP}{TP + FN}$$

4. F1 score: When the accuracy of Recall and Precision is taken, the result is F1 score, given by the following equation:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.6 Model Performance and Comparison

The performance of the three supervised models i.e. Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes has been evaluated through this study. Using an appropriate dataset all the mentioned models were trained, validated, and tested systematically. To assess the model's robustness and classification efficiency, different key performance metrics like confusion matrix, precision, F1 score and recall were used for evaluation. When the three models were compared among each other, SVM model showed a great training accuracy of 96.41%, with 74% and 55% as its respective validation and testing accuracy. Although it demonstrated a high level of precision and recall for the normal and large cell carcinoma categories, its performance weakened on squamous cell carcinoma and adenocarcinoma in the test dataset, revealing a tendency to overfit despite strong training results. In particular, the SVM obtained a validation F1-score of 0.92 for the normal class and 0.79 for large cell carcinoma, while the F1-scores on the test set significantly decreased to 0.84 and 0.47, respectively.

Logistic Regression reached a training accuracy of 96.08%, along with a validation accuracy of 78% and a test accuracy of 56%. This model showed a more balanced generalization acrossclasses compared to SVM. It maintained high recall and F1-score values for the normal class across both validation (1.00recall, 0.96 F1-score) and test sets (0.96 recall, 0.75 F1-score), suggesting robustness in identifying non-cancerous samples. However, its performance on adenocarcinoma and squamous cell carcinoma remained moderate.

The Naive Bayes classifier demonstrated the lowest accuracy during training, achieving 82.38%, while the accuracies for validation and testing were 69% and 44%, respectively.The model's simplistic probabilistic assumptions led to significant reductions in predictive performance, particularly on complex or overlapping features. For example, its test F1-score for adenocarcinoma was only 0.37, and for squamous cellcarcinoma, it was 0.36, underscoring its limited capability in handling the intricacies of HOG-based features.

SVM REPORT

Table -1: Validation Classification Report for SVM

	precision	recall	f1-score	Support
adenocarcinoma	0.70	0.61	0.65	23
Large cell carcinoma	0.88	0.71	0.79	21
Squamous cell carcinoma	0.92	0.92	0.92	13
normal	0.55	0.80	0.65	15
accuracy			0.74	72
Macro avg	0.76	0.76	0.75	72
Weighted avg	0.76	0.74	0.74	72

Table -2: Test Classification Report for SVM

	precision	recall	f1-score	Support
adenocarcinoma	0.60	0.47	0.53	120
Large cell carcinoma	0.34	0.76	0.47	51
Squamous cell carcinoma	0.77	0.93	0.84	54
normal	0.68	0.31	0.43	90
accuracy			0.55	315
Macro avg	0.60	0.62	0.57	315
Weighted avg	0.61	0.55	0.54	315

LOGISTIC REGRESSION REPORT

Table -3: Validation Classification Report for Logistic Regression

	precision	recall	f1-score	Support
adenocarcinoma	0.77	0.74	0.76	23
Large cell carcinoma	0.89	0.76	0.82	21
Squamous cell carcinoma	0.93	1.00	0.96	13
normal	0.56	0.67	0.61	15
accuracy			0.78	72
Macro avg	0.79	0.79	0.79	72
Weighted avg	0.79	0.78	0.78	72

Table -4: Test Classification Report for Logistic Regression

	precision	recall	f1-score	Support
adenocarcinoma	0.63	0.49	0.55	120
Large cell carcinoma	0.40	0.71	0.51	51
Squamous cell carcinoma	0.61	0.96	0.75	54
normal	0.64	0.33	0.44	90
accuracy			0.56	315
Macro avg	0.57	0.62	0.56	315
Weighted avg	0.59	0.56	0.55	315

NAIVE BAYES REPORT

Table -5: Validation Classification Report for Naive Bayes

	precision	recall	f1-score	support
adenocarcinoma	0.74	0.61	0.67	23
Large cell carcinoma	0.73	0.76	0.74	21
Squamous cell carcinoma	0.62	0.77	0.69	13
Normal	0.67	0.67	0.67	15
Accuracy			0.69	72
Macro avg	0.69	0.70	0.69	72
Weighted avg	0.70	0.69	0.69	72

Table -6: Test Classification Report for Naive Bayes

	precision	recall	f1-score	support
adenocarcinoma	0.68	0.25	0.37	120
Large cell carcinoma	0.42	0.75	0.54	51
Squamous cell carcinoma	0.33	0.85	0.48	54
Normal	0.57	0.27	0.36	90
Accuracy			0.44	315
Macro avg	0.50	0.53	0.44	315
Weighted avg	0.55	0.44	0.41	315

Confusion Matrix Analysis: Analysis of the confusion matrices revealed that the SVM classifier produced the fewest false positives and false negatives in the validation set, Logistic Regression showed consistent performance with fewer misclassifications in the normal class and large cellcarcinoma but struggled similarly with squamous cellcarcinoma. In contrast, the Naive Bayes model demonstrated a higher rate of misclassification across all classes during testing, indicating its relatively weak generalization capability.

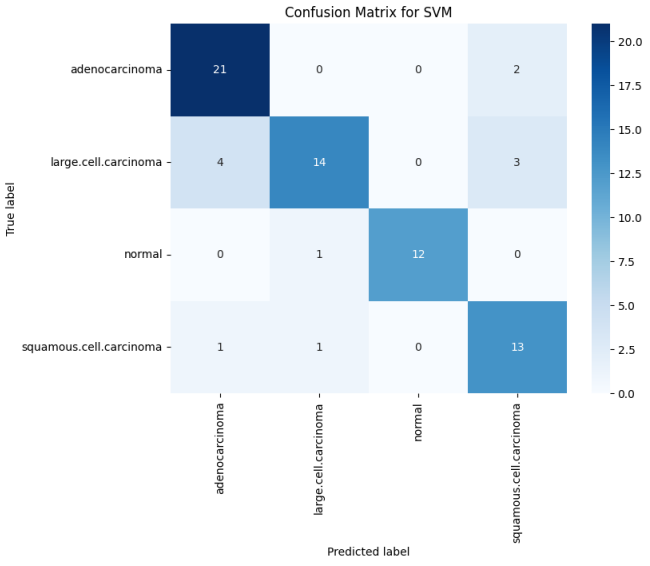

Fig -1: Validation Confusion Matrix for SVM

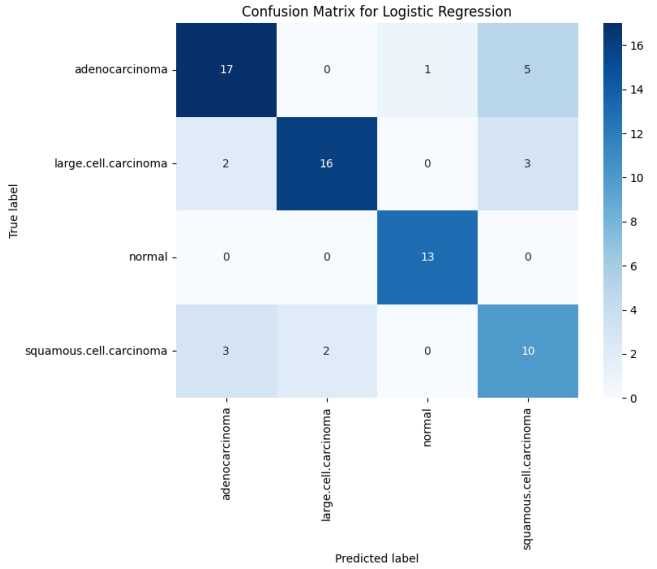

Fig -2: Validation Confusion Matrix for Logistic Regression

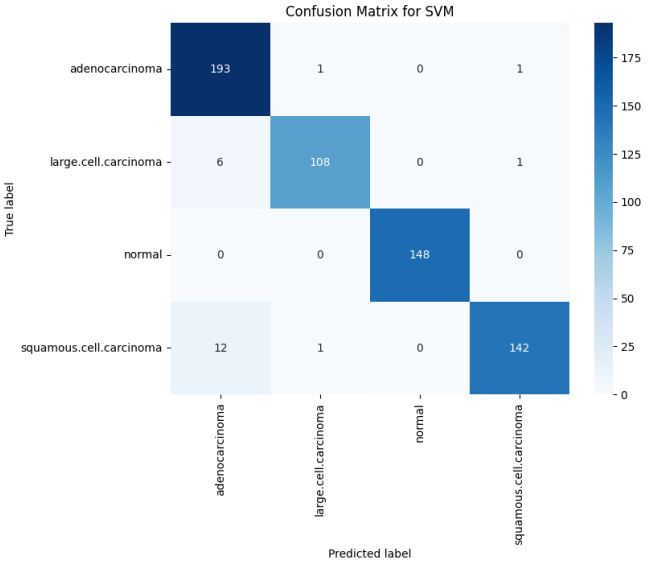

Fig -4: Training Confusion Matrix for SVM

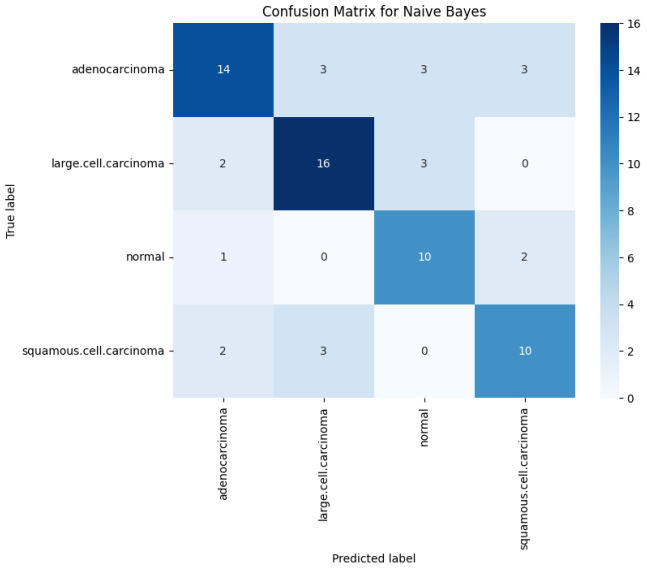

Fig -3: Validation Confusion Matrix for Naive Bayes

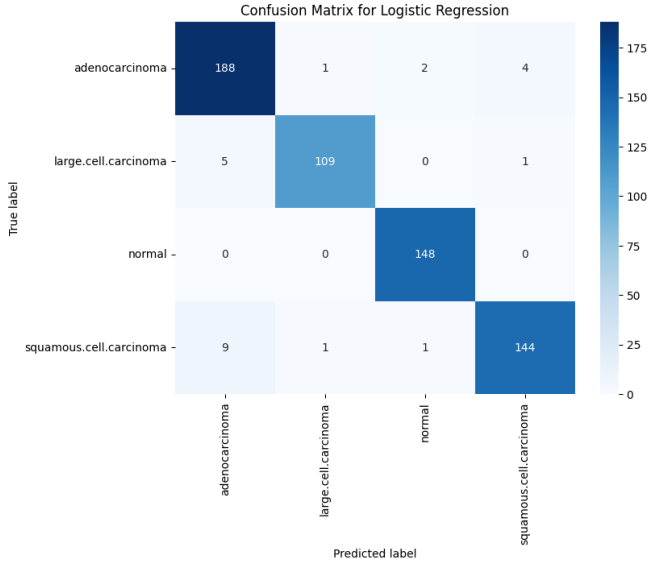

Fig -5: Training Confusion Matrix for Logistic Regression

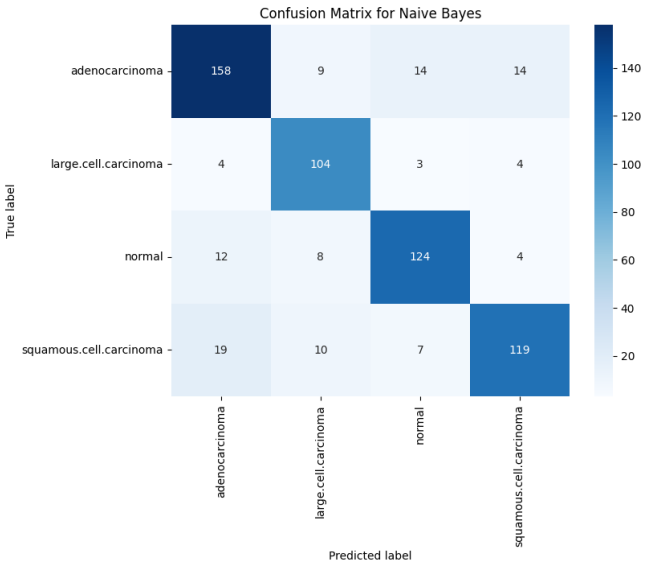

Fig -6: Validation Confusion Matrix for Naive Bayes

3.7 Tumor Localization

This research introduces an organized and modular method for identifying and measuring tumor nodules in computed tomography (CT) scans by utilizing a blend of image processing techniques and morphological analysis. The proposed methodology aims to pinpoint possible tumor areas, extract pertinent morphological features, and deliver a thorough quantification expressed in medically meaningful terms.

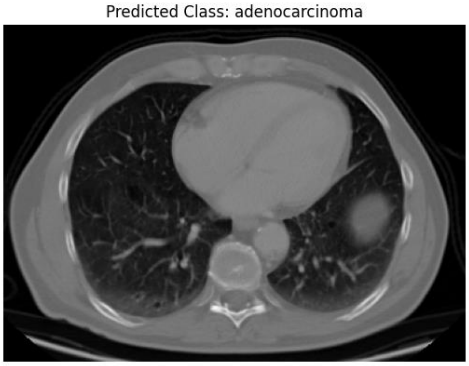

Fig -7: Input CT Scan Image

To start, CT scans are transformed to grayscale to ensure uniform processing. Segmentation of the images is done with Otsu's thresholding method, which determines the best global threshold level by considering the variance between the foreground and background level of light. This processes aids in separating bright tumor-like regions from the surrounding tissues enabling the formation of a binary image of the areas of primary focus.

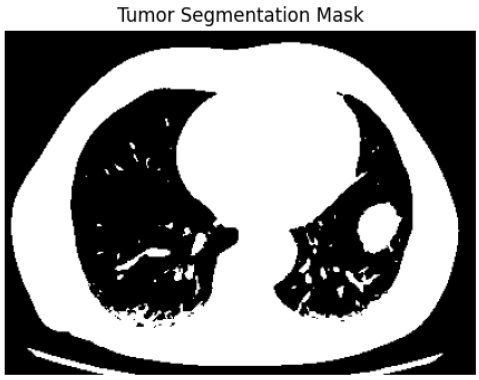

Fig -8: Tumor Segmentation Mask

Following the segmentation process, individual prospective tumors are distinguished by the segmentation of the connected blob regions. Morphological attributes of each component are evaluated, including area along with eccentricity and other shape factors, to distinguish tumor-like structures from non-relevant noisy regions. A filter is applied to retain only those regions whose shapes and complexities coincide with those expected of a tumor, largely reducing erroneous positives.

For each of the confirmed regions, some geometric values representing length, width, radius, diameter and area are captured. These values are then transformed from pixel measurements into standard medical measurement of micrometers, millimeters and centimeters with defined spatial resolution values. This information provides a quantitative assessment of tumor nodules thereby improving their clinical interpretation.

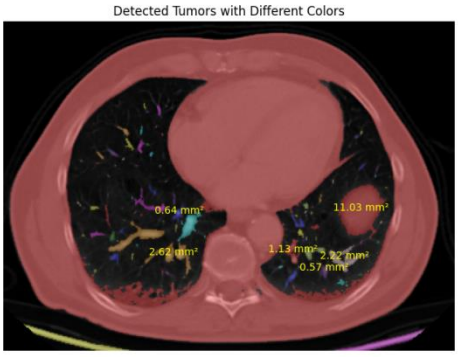

Fig -9: Tumor Detected along with its size

The entire catalog of morphological characteristics is structured to give relevant information regarding the size and sculpting of a tumor. This, apart from aiding detection, will in the future assist in diagnosis, treatment planning, and disease progression monitoring. Moreover, the features which are obtained can be used along with extraction processes to aid in classification and advanced analysis predictive analytics in imaging.

4.CONCLUSION

This study demonstrated the effectiveness of computer-based techniques on the detection of lung cancer from CT scans. While Naive Bayes was the most computationally simple to execute, SVM achieved the highest training accuracy of 96.41% followed closely by Logistic Regression with 96.08%. Apart from classification, certain segmentation and preprocessing methodologies were utilized to refine the delineation of tumor areas to ensure accuracy. The approach has the potential to improve and streamline real-world clinical decision-making and support early diagnosis. The proposed approach shows strong potential for enhancing early diagnosis and supporting clinical decision-making in real-world health caresettings.

REFERENCES

[1] A. Hazra, B. Mitra, and S. Roy, "Lung cancer survivability prediction using SVM and logistic regression," International Journal of Healthcare Information Systems and Informatics, vol. 12, no. 4, pp. 1–12, 2017.

[2] S. Agarwal, S. Thakur, and A. Chaudhary, "Performance evaluation of machine learning techniques for lung cancer prediction," International Journal of Advanced Research in Computer Science, vol. 10, no. 5, pp. 42–47, 2019.

[3] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Detection of pneumoconiosis using subset features and support vector machines," Neural Computing and Applications, vol. 19, pp. 543–550, 2010.

[4] A. Verma, R. Singh, and S. Rathore, "Comparative analysis of data mining classification algorithms for cancer prediction," Procedia Computer Science, vol. 132, pp. 400–408, 2018.

[5] M. Aljuaid, A. Almotiri, and A. Khan, "A hybrid deep learning approach for lung cancer classification using CNN and ANN," Computational Intelligence and Neuroscience, vol. 2022, Article ID 2154879, 2022.

[6] J. Lim, N. Kim, J. Seo, Y. Lee, Y. Lee, and S.-H. Kang, "Improved diagnosis of interstitial lung disease using SVM on high-resolution CT images," Medical Imaging and Diagnosis, vol. 33, no. 2, pp. 151–158, 2020.

[7] E. Avci, "A new approach for lung disease diagnosis using SVM and image texture analysis," Expert Systems with Applications, vol. 36, no. 4, pp. 7610–7615, 2009.

[8] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. Suykens, "Improved support vector machine classifiers for high-dimensional data using kernel-based learning," IEEE Transactions on Neural Networks, vol. 22, no. 6, pp. 984–995, Jun. 2011.

[9] W. A. Weber et al., "Repeatability of 18F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials," Journal of Nuclear Medicine, vol. 56, no. 8, pp. 1137–1143, 2015.