Comparison of Feature Extraction Filters For Lung Cancer Prediction

Stuti Shukla[1], Anjali Dwivedi[2], Shweta Singh[3], Naushad Ali[4], Ashutosh Singh[5]

[1,2,3,4] *Student, Department of Computer Science and Engineering, Babu Banarasi Das Institute of Technology & Management, Lucknow, India*

[5]*Assistant Professor, Department of Computer Science and Engineering, Babu Banarasi Das Institute of Technology & Management, Lucknow, India*

Abstract- Lung cancer is a deadly form of cancer that is difficult to detect. It is more important for care to inspect nodules soon and correctly since it generally causes death in both men and women. The number of people diagnosed with lung cancer is directly proportionate to the number of chain smokers. As a result, a variety of procedures have been developed to identify lung cancer in its early stages. A comparative analysis of multiple machine learning-based approaches for lung cancer diagnosis has been reported in this progress. In addition, to use image recognition to identify lung cancer nodules, many classifier methods are linked with numerous segmentation algorithms. [1] CT scan images have been discovered to be more acceptable for having reliable results in this investigation. The experiment is divided into two parts: Identify the most important feature used in lung cancer analysis by CT scan and map it to a computer-related format in the first step. In the second phase, machine learning techniques are used to pick and extract features. As a result, CT scan images are commonly utilized to diagnose cancer. The classification method, Random Forest Classifier was used to investigate lung cancer prediction. The results reveal that classification accuracy improves in the vast majority of situations, indirectly proving reliability. For feature extraction, we used three filters – Gabor, Local Binary Pattern, and Histogram of Oriented Gradients. Hence, their comparisons are shown in this paper.

Keywords- Gabor, Histogram of Oriented Gradient, Local Binary Patterns, Lung Cancer Prediction, Random Forest.

I. INTRODUCTION

Cancer is one of the deadliest diseases known to mankind. Because of late diagnosis, it has the potential to kill people all around the world. The most fundamental function of the lungs is to provide oxygen to the body and remove carbon dioxide during vital activities. Lung cancer develops when tissues and cells in the lungs proliferate uncontrollably. When these masses expand out of control in their surroundings, they can cause injury to the surrounding tissues. Lung cancer is the most common cause of mortality in men and the second most common cause of death in women. Lung cancer claims the lives of over 1.3 million individuals per year around the world. Lung cancer affects 30-40 thousand persons in Turkey, every year.

Lung cancer symptoms may not manifest themselves until the disease has progressed significantly. The most crucial aspect that makes lung cancer so dangerous is that it progresses without symptoms. Almost a quarter of cancer patients show no symptoms. The majority of people learn that lung cancer is caused by X-rays taken as a result of another illness. In the case of lung cancer, early detection is crucial. Lung cancer has a high proclivity for spreading to the bones, liver, brain, and adrenal glands. The average life expectancy and quality of life have improved as a result of newly developed lung cancer treatment approaches. Lung cancer can now be discovered at an early stage because of developments in imaging techniques such as low-dose spiral computed tomography.

This process may be made considerably more efficient by using image processing techniques and machine learning. As digital technologies grow more integrated into our daily lives, they may become an important part of medical diagnostics. Lung cancer is the most frequent type of cancer among men, both in terms of occurrence and fatality. It is the third most common cancer in women and has the second-highest fatality rate after breast cancer. Lung cancer has the greatest fatality rate of any type of cancer. It also has the lowest post-diagnosis survival rate of all forms of malignancies, with an annual increase in the number of deaths. Every year, over 160,000 people die from malignant cancer, which affects around 2,20,000 people. Survival from lung cancer is directly related to its growth at its detection time, so the chances of survival increase if the cancer is detected in the early stages. The main contributor to lung cancer is smoking. An estimated 85 percent of lung cancer cases in males and 75 percent in females are caused by cigarette smoking. Other causes include radon gas, asbestos, air pollution, genetics, etc. The population segment most likely to develop lung cancer is people aged over 50 who have a history of smoking.

CT scans are reported to be more successful than ordinary chest X-rays in detecting and diagnosing lung cancer. Unlike typical X-rays, which only show dense body parts like bones, CT images show the soft tissues of the body, including blood arteries, muscle tissue, and organs like the lungs. CT pictures show cross-sections of the body, whereas traditional X-rays show flat 2D images. A typical CT scan image is depicted in Figure 1.

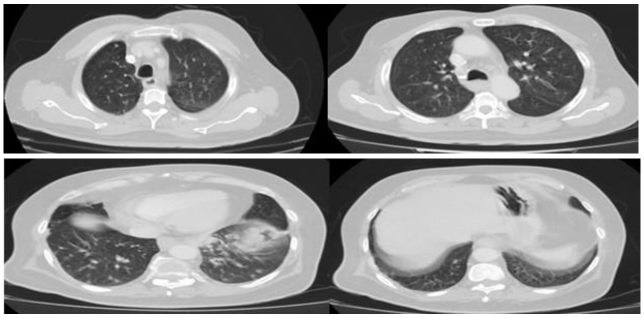

Figure 1. Images from CT scans were acquired at various time slices.

The goal of this study is to use the CT scan data to classify lung cancer in patients. Using Gabor, LBP, and HOG feature extraction filters, as well as a Learning algorithm. The comparative examination of the various feature extraction filters will reveal that raising the diagnostic procedure's efficiency and accuracy reduces the massive false positive rate. In Section I of this paper, we discuss lung cancer and related lung cancer detection efforts.

The rest of the paper is organized as follows: the related work is presented in the following section. The methodology for developing experiments is defined in section 3. The findings of the experiments are provided in section 4. Finally, in section 5, the conclusions are drawn followed by the future scope of this model.

II. RELATED WORK

As cancer is one of the most lethal diseases in the world, many researchers have focused on one of the most important cancer biomarkers, DNA methylation, and used feature selection or feature extraction techniques on its massive data to improve prediction accuracy by obtaining the best features set that discriminate biological samples of various cancer types.

The study reported by Wu et al. [3] is an example of feature selection strategies being used in research. This work used a three-step feature selection method based on the properties of clinical DNA methylation data, and the feature selection procedure chose numerous cancer-related and lymph node metastasis-related gene biomarkers. The results of the approach showed that the accuracy of prediction in detecting LN metastasis was greatly enhanced.

Furthermore, employing probe-level DNA methylation data, Baur et al. [4] developed a unique feature selection technique based on sequential forward selection to compute gene-centric DNA methylation. The results of the proposed approach using the K-Nearest Neighbors classification outperformed other algorithms on all metrics, and it was able to accurately predict the expression of specific genes using just DNA methylation data. The findings also revealed that those DNA methylation-sensitive genes were overrepresented in Gene Ontology terms related to biological process regulation.

By picking only the useful features from the whole feature set, Kaur et al. [5] validated the importance of using feature selection to forecast diseases such as breast cancer, lung cancer, and heart disease. The study compared the accuracy and efficiency of various feature selection approaches, including F-score, Genetic Algorithm, K-means, and SVM-RFE, and found that SVM-RFE attained the greatest accuracy of 97 percent utilizing the support vector machine (SVM).

Singh et al. [6] also presented a review that highlighted the importance of the feature selection algorithm in improving classifier accuracy. This analysis revealed that each feature selection algorithm behaves differently and has its own set of benefits and drawbacks. The study found that the feature selection method is a crucial factor in the wrong categorization of huge data, as it selects essential features while ignoring irrelevant ones, resulting in a decrease in classification accuracy.

[7] Discovered a hybridization of the feature selection elimination approach and a machine learning algorithm based on Random Forrest. The goal of this work was to develop a two-stage computer-aided diagnostic system to detect benign breast tumors from malignant ones, with the first stage of the proposed system performing a data reduction procedure in preparation for the second stage's learning algorithm. The proposed approach of this study beat other studies in the test phase, with a classification accuracy of 99.82 percent and 99.70 percent, respectively. Recent studies, on the other hand, focused solely on feature extraction approaches in order to speed up and enhance prediction accuracy.

Using a fuzzy interference system and an active contour model, Roy, Sirohi, and others [8] created a system to detect lung cancer nodules. For visual contrast enhancement, this method employs grey transformation. Before segmentation, the image is binarized, then the resulting image is segmented using an active contour model. The fuzzy inference method is used to classify cancers. To train the classifier, features such as area, mean, entropy, correlation, main axis length, and minor axis length are extracted. The system's overall accuracy is 94.12 percent. Among its limitations, it does not identify cancer as benign or malignant, which is the suggested model's future scope.

Watershed segmentation was used by Ignatious and Joseph [9] to create a system. It uses the Gabor filter to improve image quality during pre-processing. The accuracy of the neural fuzzy model and the region growing method are compared. The suggested model has a 90.1 percent accuracy, which is greater than the model with segmentation using a neural fuzzy model and the region growing approach. This model has the advantage of using marker-controlled watershed segmentation, which eliminates the problem of over-segmentation. As a drawback, it does not distinguish between benign and malignant cancers, and while accuracy is good, it is still not sufficient. Some modifications and contributions to this model have the potential to improve the accuracy to a desirable level.

III. METHODOLOGY

A. Data and Pre-processing

The Japanese Society of Radiological Technology (JSRT) [14] provided the Standard Digital Image Database for this investigation. The dataset as shown in Fig. 2 contains 1000 CT (computed tomography) pictures. The photos are labeled with the presence or absence of a lung nodule. The nodule status of half of the pictures is determined, while the nodule status of the other half is determined. Images are saved as a 1024*1024 matrix with 1 byte of data per row. For classification, the dataset is separated into two parts: a train set and a test set. We randomly select 80 percent of the data as a train set and 20 percent of the data as a test set from the dataset.

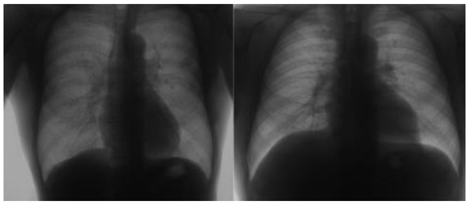

Fig. 2. The computed tomography pictures (left image with nodule, right image without nodule).

B. *Feature Extraction*

Filters used for feature extraction are explained below:

1) *Local Binary Pattern*

Local Binary Pattern (LBP) is a basic yet effective texture operator that labels pixels in an image by thresholding each pixel's neighborhood and treating the result as a binary number. The LBP texture operator has become a prominent method in a variety of applications due to its discriminative power and computational simplicity. It can be viewed as a unifying approach to texture analysis's typically diverse statistical and structural concepts. The LBP operator's resistance to monotonic gray-scale changes produced, for example, by illumination variations is perhaps its most essential quality in real-world applications. Another key feature is its computational simplicity, which allows it to evaluate photos in difficult real-time scenarios. [10]

2) *Histogram of Oriented Gradients*

The Histogram of Oriented Gradients (HOG) is a feature descriptor used for object detection in computer vision and image processing. The technique counts the number of times a gradient orientation appears in a certain area of an image. Edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts are all comparable methods, but this one differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for enhanced accuracy. Because of this normalization, the invariance to changes in illumination and shadowing is improved.

Compared to other descriptors, the HOG descriptor has a few major advantages. Except for object orientation, it is invariant to geometric and photometric alterations because it operates on local cells. Such changes would only be visible in wider geographic areas.[11]

3) *Gabor*

A Gabor filter, named after Dennis Gabor, is a linear texture analysis filter in image processing. It examines if the image contains any specific frequency content in specific directions in a confined region around the point or region of study. Many modern vision experts say that Gabor filter frequency and orientation representations are similar to those of the human visual system. They've been discovered to be especially useful for texture representation and discrimination. A 2-D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave in the spatial domain.

The text is rich in high-frequency components, whereas pictures are largely smooth in nature, Gabor filters with varying frequencies and orientations in different directions have been used to localize and extract text-only patches from complicated document images (both grey and color). Gabor filters are also commonly employed in pattern recognition software.[12][13]

C. *Classification Algorithm*

- *Random Forest*

Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression problems. It creates decision trees from various samples, using the majority vote for classification and the average for regression. One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification.[14] For classification difficulties, it produces superior results.

The flow chart of the proposed model is shown below in Fig 3.

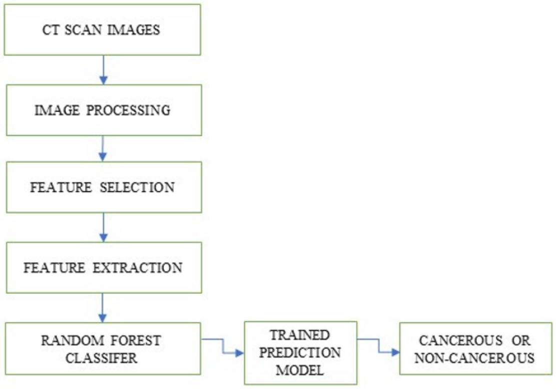

Fig. 3. Proposed Model

IV. EXPERIMENTAL RESULTS

Results obtained from the above-mentioned filters by applying the machine learning technique are mentioned in the below table.

Table I
Comparative analysis of feature extraction filters

S.NO.	FILTER	ACCURACY
1	LBP	91.77%
2	HOG	85.01%
3	GABOR	96.00%

The Confusion Matrix and the ROC Curve of the respective filters were also obtained while doing the experimental analysis.

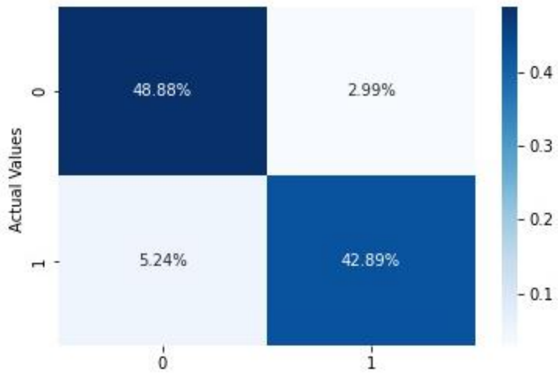

Fig. 4. The confusion matrix was obtained from the LBP filter.

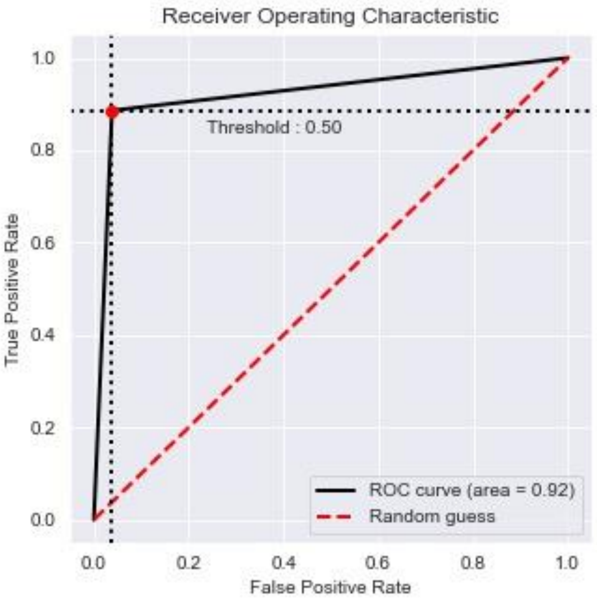

Fig. 5. The ROC Curve was obtained from the LBP filter.

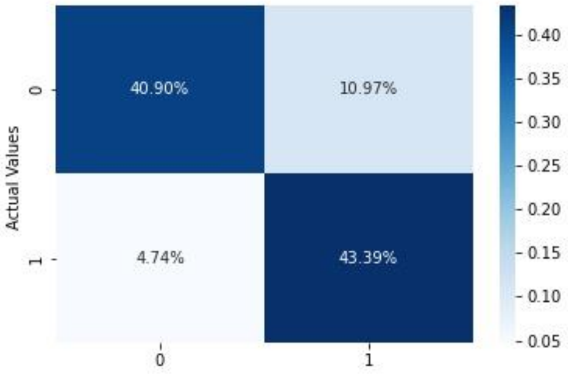

Fig. 6. The confusion matrix was obtained from the HOG filter.

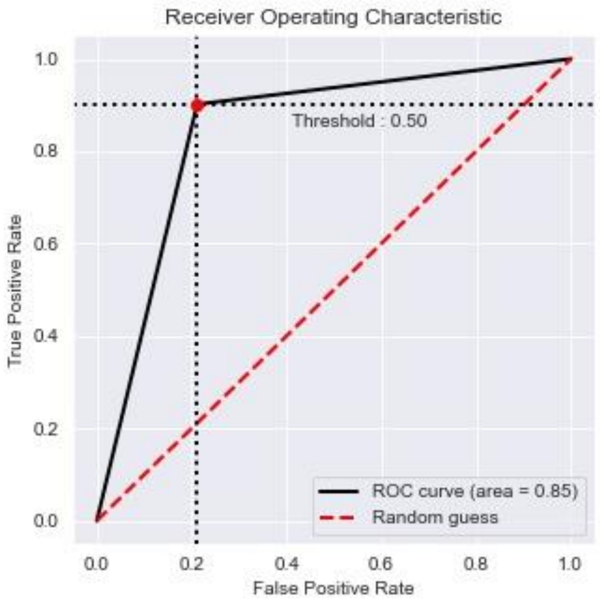

Fig. 7. The ROC Curve was obtained from the HOG filter.

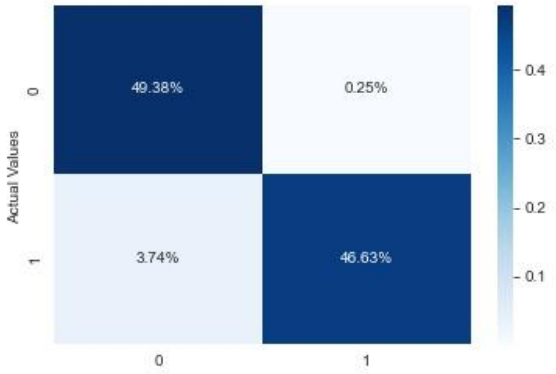

Fig. 8. The confusion matrix was obtained from the Gabor filter.

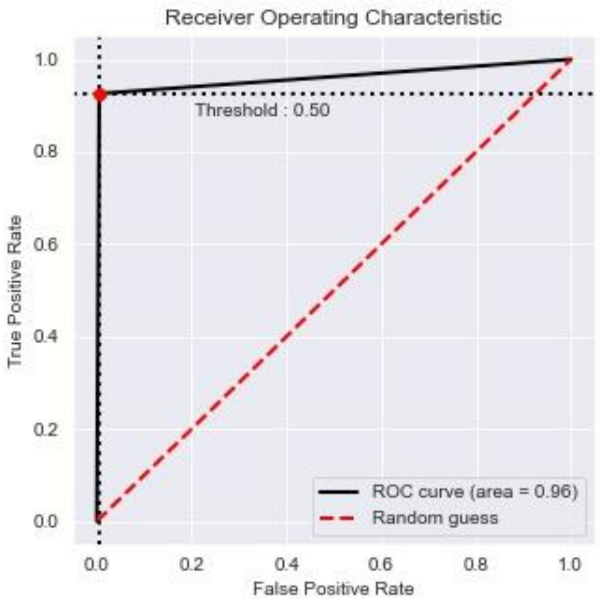

Fig. 9. The ROC Curve was obtained from the Gabor filter.

V. CONCLUSION AND FUTURE SCOPE

We used various feature extraction approaches to detect lung cancer from chest CT scans in this investigation. In our situation, it did not result in a significant loss of data. Accuracy results had a minor impact on small amounts that might be overlooked to save time and space. Previously, a doctor had to perform many tests to determine whether or not a patient had lung cancer, which was a time-consuming process. The machine learning algorithm employed in this comparison study was the Random Forest classifier. The degrees of accuracy demonstrated by these proposed methods are impressive. As the dataset grows, the accuracy will deteriorate. The present best model does not produce good accuracy results and does not classify the degree of malignancy in discovered nodules. As a result, a new system is suggested. Feature extraction techniques are used to detect the malignant nodule from a lung CT scan image in the suggested system. The suggested approach diagnoses cancer more accurately than the current model, and the Gabor filter-based classifier has a 96 percent accuracy. In comparison to the existing best model, the suggested system shows overall improvement. As a result, future scope enhancement in this area can be accomplished by implementing classification in stages. Furthermore, proper pre-processing and the elimination of spurious objects can improve accuracy even more.

REFERENCES

[1] Makaju, S., Prasad, P.W.C., Alsadoon, A., Singh, A.K., Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125: 107-114.

[2] John N. Korecki; Yoganand Balagurunathan; Yuhua Gu; Virendra Kumar Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features,2014.

[3] J. Wu, Y. Xiao, C. Xia, F. Yang, H. Li, Z. Shao, Z. Lin, and X. Zhao," Identification of Biomarkers for Predicting Lymph Node Metastasis of Stomach Cancer Using Clinical DNA Methylation Data," Disease Markers, 2017.

[4] B. Baur and S. Bozdag, "A Feature Selection Algorithm to Compute Gene Centric Methylation from Probe Level Methylation Data," PLoS One, February 2016.

[5] S. Kaur and S. Kalra," Feature Extraction Techniques using Support Vector Machines in Disease Prediction," IJARSE, vol. 5, May 2016.

[6] R. Singh and M. Sivabalakrishnan, "Feature Selection of Gene Expression Data for Cancer Classification: A Review," Procedia Computer Science, vol. 50, pp. 52-57, January 2015.

[7] C. Nguyen, Y. Wang, and H. Nguyen, "Random Forest classifier combined with feature selection for breast cancer diagnosis and prognostic," Journal of Biomedical Science and Engineering, vol. 6, pp. 551-560, 2013.

[8] Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation. DOI: 10.1109/CCAA.2015.7148560.

[9] Ignatious, S., & Joseph, R. (2015) "Computer-aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.

[10] Bin Xiao; Kaili Wang; Xiuli Bi; Weisheng Li; Junwei Han; An Enhanced Local Binary Feature for Texture Image Classification,2019.

[11] Zarif Al Sadeque; Tanvirul Islam Khan; Qazi Delwar Hossain; Mahbuba Yesmin Turaba, Automated Detection and Classification of Liver Cancer from CT Images using HOG-SVM model, 2019.

[12] Büşranur Bahat; Pelin Görgel; Lung Cancer Diagnosis via Gabor Filters and Convolutional Neural Networks, 2021

[13] Mohammad A. Alzubaidi; Mwaffaq Otoom; Hamza Jaradat Comprehensive and Comparative Global and Local Feature Extraction Framework for Lung Cancer Detection Using CT Scan Images, 2021.

[14] J. Alam, S. Alam, and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classified," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.