Leveraging Correlation Analysis to Enhance Lung Cancer Detection with KNN and CNN

Eshaan Sowale
Electronics and Telecommunication
Symbiosis Institute Of Technology
Symbiosis International (Deemed University)
Pune, India

Mitali Pagarware
Electronics and Telecommunication
Symbiosis Institute Of Technology
Symbiosis International (Deemed University)
Pune, India

Azhan Farooqui
Electronics and Telecommunication
Symbiosis Institute Of Technology
Symbiosis International (Deemed University)
Pune, India

Mohammad Daniyal Khan
Electronics and Telecommunication
Symbiosis Institute Of Technology
Symbiosis International (Deemed University)
Pune, India

Anurag Mahajan
Electronics and Telecommunication
Symbiosis Institute Of Technology
Symbiosis International (Deemed University)
Pune, India

***Abstract*—Lung Cancer is the type of cancer that begins in the lungs. Over time there is an increase in seriousness. But the symptoms are visible only when the disease is advanced. Hence, it is very crucial to detect it at an early stage . There are several methods for detection using Deep Learning, Machine Learning and Convolution Neural Network. Here in this study K-Nearest Neighbors (KNN) algorithm is used for Image Processing of images and implemented using Python. The images are compared with the data set that has both the cancer affected and unaffected images. The input images are compared to the images in the data set and whether the cancer is presence of cancer is found. Image Processing technique is used to manipulate images using algorithms and several techniques. It is used to enhance the images, improve the quality and remove noise. It involves Image Analysis, Image Compression, Image Restoration, Image Filtering and Image Synthesis.**

***Index Terms*—Cancer Detection,K-Nearest Neighbour, Image Processing, Image Segmentation,Feature Extraction**

I. Introduction

The chest contains lungs that are two sponge like structures for respiration. It maintains the oxygen level of the human body. The air that we breathe or inhale through nose or mouth goes into the lungs through windpipe. Finally, the exchange of gases occur in the sac like structure known as alveoli. Lung cancer starts in the cells lining the alveoli. Active Smoking and Passive Smoking are the main reasons for lung cancer. The former includes tobacco smoke and the latter includes exposure to smoking. There are various other factors as well that contribute to lung cancer such as exposure to particulate matter and carcinogen. Chest pain, shortness of breath, blood while coughing, and fatigue are some of the consequences.

The overall health of the patient, the type of cancer and its stage determine the type of treatment required. Chemotherapy kills or prevents the rapid growth of cancer cells in the body, shrinks tumors and prevent spread of cancer cells and to extend the survival of patient. On the other hand, Radiotherapy uses high-energy radiation such as X-Rays and Gamma Rays to destroy the cancer cells. The side effect of such therapy is that they also affect healthy cells in human body leading to decreased immunity. Hence, early detection is important so that proper treatment can be received depending on the type of cancer. Computerized tomography (CT) uses computer processing to make cross sectional images of soft tissues inside body. But still it is difficult for doctors to detect cancer from CT scan images. Images must go through pre-processing stages for better quality. Hence computer techniques are used to do so. The aim of this project is to study the technique and implement it and analyse the image using digital image processing. Lung has two categories: Small cell and non-small cell.

II. Types Of Lung Cancer

Mainly there are two types of Lung Cancer :

A. Non-small Cell :

They constitute four out of five of lung cancers which is around Eighty percent. It majorly occurs in people who actively smoke but is also commonly found in people who don't smoke.

B. Small Cell :

The remaining one tenth are the small cell lung cancer. It is found to spread faster than the other type of cancers. Hence, chemotherapy and radiation therapy are used. Most in most of the cases the cancer returns at some point.

III. Methodology

A. Data Acquisition and Pre-processing

1) Image Data set Collection: Obtain a data set of lung images. It must have images of several cases. The data set should ideally consist of high-resolution medical images such as CT scans or X-rays. Collaborate with healthcare institutions or research organizations to ensure access to appropriate data.

2) Data Pre-processing: Convert the images to a standardized format (e.g., DICOM).

Apply image pre-processing techniques to enhance image quality, such as noise reduction, contrast adjustment, and image resizing. Extract region of interest (ROI) if necessary, focusing on the lung area to reduce computational complexity and improve model accuracy. Normalize pixel values to ensure uniformity across images.

B. Feature Extraction

1) Feature Engineering: Utilize digital image processing techniques to extract meaningful features from the pre-processed images. Common features may include Texture features (e.g., Haralick, Gabor, or LBP features) to capture textural patterns. Shape features (e.g., contour-based features) to describe lung shape characteristics. Intensity-based features (e.g., histogram statistics) to capture pixel intensity information.

2) Feature Selection: Evaluate the extracted features and perform feature selection to retain the most relevant ones. Techniques like feature ranking or dimensionality reduction (e.g., PCA) can be applied to optimize feature sets.

C. Data Splitting

1) Training and Testing Set Creation: Split the data set .Ensure that an appropriate balance between cancerous and non-cancerous images is maintained in both sets to prevent class imbalance issues.

D. KNN Model Development

1) KNN Classifier Implementation : Implement the K-Nearest Neighbors (KNN) algorithm using a suitable machine learning library such as scikit-learn.

2) Hyper-parameter Tuning: : Determine the optimal value of 'k' (number of neighbors) through cross-validation techniques (e.g., grid search) on the training data.

3) Model Training: Train the KNN model using the training set and the selected hyper-parameters.

E. Model Evaluation

1) Performance Metrics: Evaluate the KNN model's performance on the testing data set using various performance metrics.

2) Confusion Matrix: In order to check the ability of the model visualize the confusion matrix to detect cancerous and non-cancerous lung images.

F. Interpretation and Visualization

1) Feature Importance: Analyze the importance of the extracted features in making predictions and visualize them using appropriate visualization techniques.

2) Result Visualization: Create visualizations, such as ROC curves or precision-recall curves, to illustrate the model's performance and effectiveness.

IV. Image Enhancement

It is the process of enhancing the quality of the image example improving them visually, enlarging or bringing out special features, removing noises etc.

Image Enhancement is classified into two categories : **Spatial Domain** and **frequency Domain.**

V. Median Filter

Also known as non-linear filtering, it is a type of filter used to get rid of salt and pepper noise which includes removing pixels with extreme values while preserving edges of digital images. It is also used to sharpen the contrast. .It works by replacing each pixel's value with the median value of the pixels in its neighborhood. "Window" is the size of the neighbourhood which is a square or rectangular region.It is centered around the pixel that is being processed. They are preferably used when other filters such as mean filter that replace pixel value with average value of neighbourhood cause blur or distortion in image. The limitation of a median filter is that it is not effective in removing Gaussian Noise as it is smooth and does not have extreme values. In such cases other filters are used. Following is the working of median filter:

1. To define the size of the neighbourhood, the size is usually an odd number
2. Overlap the neighbourhood window over the pixel to be filtered
3. Arrange pixel values in ascending order with the neighbourhood values
4. Replace the value of the pixel with the median value

VI. Marker Controlled Watershed Segmentation

The segmentation techniques based on edge and region are part of the segmentation approach. One of these techniques is Thresholding. An image that has pixels in two values, 1 and 0, is called a binary image. The advantage of using Threshold Segmentation is that it requires less storage space and has faster speed compared to grayscale images.

Marker-Controlled Watershed Management is employed when other techniques, such as Thresholding and Edge Detection, do not yield considerable results. It offers several advantages over other methods. It can be used to segment irregular shapes, and the user has manual control over marker placement. It provides high accuracy, one-pixel-wide, closed, and exact location of the outline, with transformations in grayscale.

It is mandatory to divide the image into segments, using a binary image. The grayscale image is first converted into a binary image. It employs the concept of local minima of the image, treating the image as a surface. Markers represent regions of interest, involving:

1) Watershed Transformation: It is used to segment an image based on the topography of the image intensity. The intensity values are represented as elevations, where bright regions are considered as high elevations, and dark levels are considered as low elevations.

2) Gradient Image: This is used to define the topography of the image, serving as the starting step for the segmentation process. Intensity transitions are highlighted in the original image.
3) Watershed Lines: These act as segmentation boundaries between regions.
4) Merge and Refinement: To avoid extra regions generated during segmentation, post-processing steps are followed to merge small regions that are not areas of interest.

VII. Feature Extraction

It is the process of obtaining numeric values from given raw data and extracting useful characteristics that are important for image analysis. These features help generate patterns within an image. To obtain a broader range of image data, several feature extraction methods can be used together. The original information from the data set is preserved.

Applications of feature extraction include Edge Detection, Texture Analysis, Segmentation Features, etc.

VIII. K-Nearest Neighbour

K-Nearest Neighbors (KNN) is an algorithm that groups data based on similarity to already existing examples. Grouping is determined by the neighbors of a data point. KNN is dependent on the similarity of features, and it is abbreviated as KNN. Selecting the right value of K is important for accuracy.

IX. Literature Review

· The data set from PRISM Diagnostics Centre, Solapur, Maharashtra, was used by Dhaware in 2016 in his study that classified lung city images as normal or abnormal. The study showed higher accuracy than other methods. However, the elimination of features was not successful using the Feature Subset Matrix, and very limited features were used for classification. The accuracy and sensitivity were not measured.

Total three segmentation methods were studied in terms of pre-processing time taken for segmentation by Bariqi Abdillah Alhadi Bustamam in 2016 by collecting a data set from IMBA Home. To classify whether the lung is normal or cancerous, a threshold value of 17178.48 was used. The drawback was that the performance of this method was not evaluated in terms of the measured parameters.

The problem of less accuracy and a limited data set was faced by S. Kalaivani and Pramit Chatterjee in 2017, who used The Cancer Imaging Archive (TCIA). The efficiency was found to be 78 percent, which could be further increased by using an efficient filtering technique and feature extraction method.

The study by Kanitkar and Sayali Satish in 2015 did not involve the use of any feature extraction technique. The use of Marker-controlled Watershed for segmentation gave an accuracy of 92 percent. The use of a better algorithm and advanced extraction process could increase the accuracy.

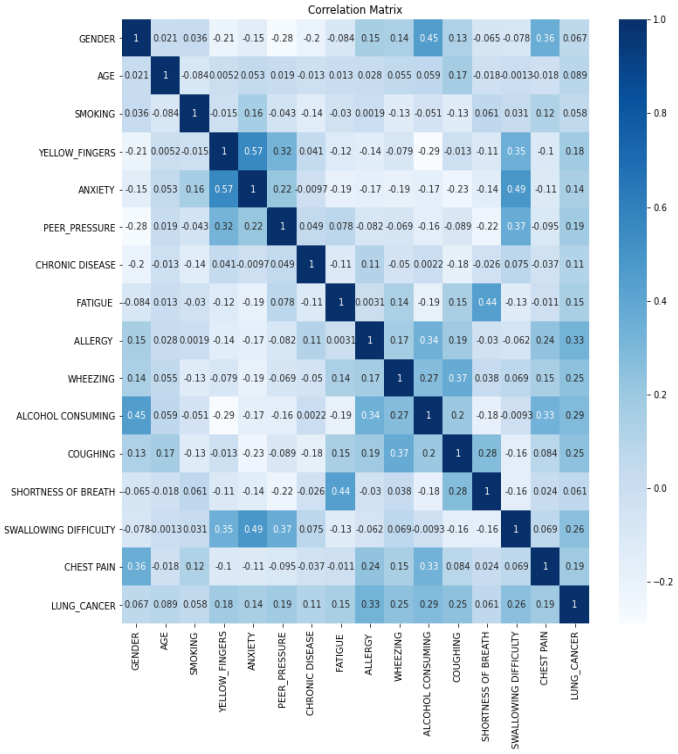
The Applied Automatic Region Growing (ARG) algorithm was used for segmentation by Manikandan Devi B in 2019 with a data set from Bharat Scans, Chennai. The results were: 94 percent accuracy, 100 percent sensitivity, and 93 percent specificity. But the data set was very limited. The accuracy can be further increased by using a suitable algorithm.

An accuracy of 78.95 percent, precision of 0.77 percent, and recall of 0.83 were observed by Islam et al., 2019, using cancerimagingarchive.net. But the data set was very limited. The accuracy was increased by using a parametric approach. The accuracy can be increased more by using the best segmentation techniques.

Various lung cancer stages like I, II, III, and IV cancers could be detected in the study by Jouy et al., 2019. The data set IMBA Home (VIAELCAP Public Access) was used, and the success rate was recorded to be 95.32 percent. The drawback was that there was no filtering technique used to remove noise. Pre-processing methods can be used to increase the performance of the system.

X. Results

Fig. 1. Correlation Matrix Output



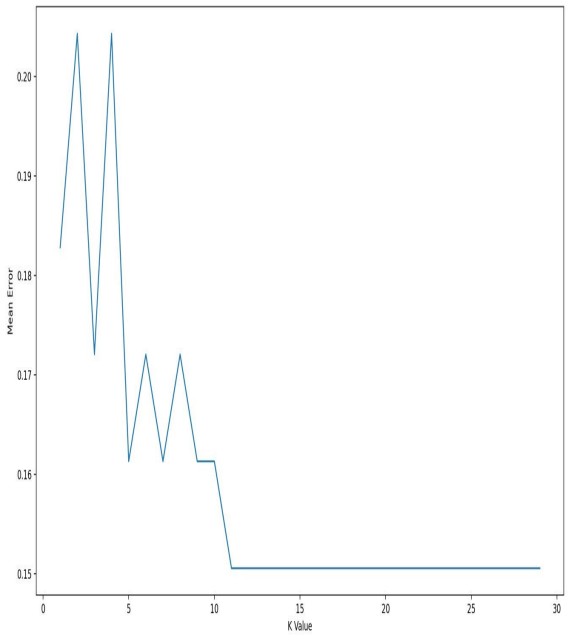
The first image is a heatmap of the lung cancer dataset's relationship matrix. The heatmap depicts the relationship between each feature pair in the dataset. A correlation coefficient of 1 indicates no correlation, a coefficient of -1 suggests that there is a perfect negative correlation, and a correlation factor of 0 means that there is no linear correlation.

The heatmap shows a proportional relationship between factors such as smoking and gender features to lung cancer. The smoking parameter also proportionately corresponds to other health-related features such as wheezing, coughing, and shortness of breath. The heatmap directly correlates different

features with each other to identify which are the most related to lung cancer. These results can be used to improve and strategize tests for lung cancer.

.

Fig. 2. Mean Error VS K-Value



The second graph is a plot of the mean error of the K-Nearest Neighbors (KNN) classifier for various values of the 'n_neighbor' parameter. Overall, the plot illustrates that the KNN classifier can achieve a high accuracy score even with a small number of neighbors. This is significant because employing a small number of neighbors speeds up and improves the efficiency of the classifier.

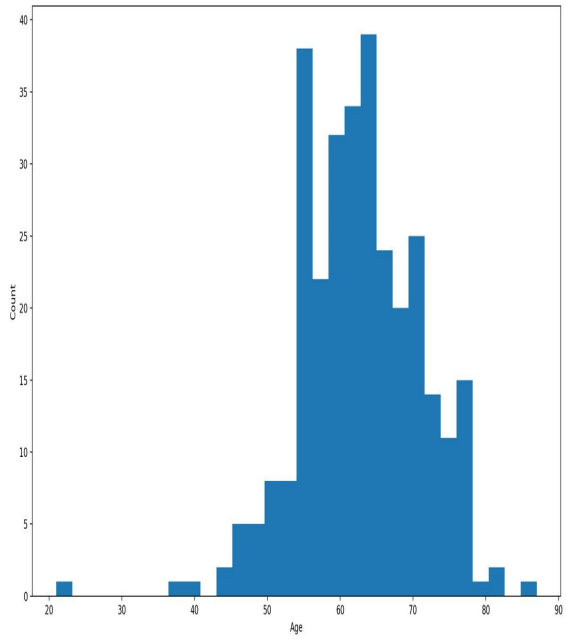
The output image is a confusion matrix generated by a K-Nearest Neighbors (KNN) classifier on the lung cancer dataset's test set. The confusion matrix displays the number of instances for each class that were categorized correctly and inaccurately. Overall, the confusion matrix provides a good summary of the KNN classifier's performance on the test set. The classifier has a high accuracy rating; however, it does make some errors. The classifier's false negative and false positive rates can be decreased by using additional data to train the algorithm for classification or by utilizing various machine learning techniques.

According to the confusion matrix, the KNN classifier correctly categorized 44 of 47 LUNG_CANCER occurrences and 28 of 33 NO_LUNG_CANCER instances. The classifier's accuracy score is 91 percent, which is a decent result.

XI. Conclusion

In this study, we introduced a K-Nearest Neighbors (KNN) classifier for lung cancer diagnosis. The KNN classifier is a simple yet effective machine-learning process that can be utilized for classification. On a lung cancer dataset, we assessed the performance of the KNN classifier. The dataset contains 80 instances, 47 of which belonged to the LUNG_CANCER class and 33 to the NO_LUNG_CANCER class. For training and testing, we employed a 70-percent/30-percent split of the dataset. On the test set, the KNN classifier had an accuracy score of 91 percent. Overall, the KNN classifier performed well on the lung cancer dataset, correctly classifying over 90 percent of the cases in the test set. The classifier can be utilized to create an accurate and efficient lung cancer diagnosis test. To improve future models of KNN, we need to collect more data to train the classifier. Other machine learning algorithms such as SVM could be compared to the KNN classifier to identify the best classifier for lung cancer detection.

Fig. 3. Count VS Age



References

[1] Dhaware, B., and Pise, A. (2016). Lung cancer detection using bayasein classifier and FCM segmentation. In 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) (pp. 170–174).

[2] Dhaware, B., and Pise, A. (2016). Lung Cancer Detection System by Using Bayesian Classifier. JournalNX, 1–5.

[3] Patel, T., and Nayak, V. (2018). Hybrid approach for feature extraction of lung cancer detection. In 2018 Second international conference on inventive communication and computational technologies (ICICCT) (pp. 1431–1433).

[4] Lobo, P., and Guruprasad, S. (2018). Classification and segmentation techniques for detection of lung cancer from CT images. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1014–1019).

[5] Asuntha, A., and Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications, 79, 7731–7762.
[6] Foeady, A., Riqmawatin, S., and Novitasari, D. (2021). Lung cancer classification based on CT scan image by applying FCM segmentation and neural network technique. TELKOMNIKA (Telecommunication Computing Electronics and Control), 19(4), 1284–1290.
[7] Bhardwaj, M. T. S. S. (2019). Lung Cancer Detection Using D and Artificial Neura.
[8] Bharathy, S., Pavithra, R., and others (2022). Lung Cancer Detection using Machine Learning. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 539–543).
[9] Lalitha, S. (2021). An automated lung cancer detection system based on machine learning algorithm. Journal of Intelligent and Fuzzy Systems, 40(4), 6355–6364.
[10] Paliwal, G., and Kurmi, U. (2021). A Comprehensive Analysis of Identifying Lung Cancer via Different Machine Learning Approach. In 2021 10th International Conference on System Modeling and Advancement in Research Trends (SMART) (pp. 691–696).
[11] Zarin Anjuman Sejuti, and Md Saiful Islam (2023). A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation. Sensors International, 4, 100229.
[12] M. N. Ab Wahab, A. Nazir, A. T. Zhen Ren, M. H. Mohd Noor, M. F. Akbar and A. S. A. Mohamed, "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi," in IEEE Access, vol. 9, pp. 134065-134080, 2021, doi: 10.1109/ACCESS.2021.3113337.
[13] B., Srinivas and Rao, Gottapu. (2019). A hybrid CNN-KNN model for MRI brain tumor classification. International Journal of Recent Technology and Engineering. 127. 20-25. 10.35940/ijrte.B1051.078219.
[14] Fatemeh Sharifonnasabi, Noor Zaman Jhanjhi, Jacob John, Peyman Obeidy, Shahab S. Band, Hamid Alinejad-Rokny, and Mohammed Baz (2022). Hybrid HCNN-KNN Model Enhances Age Estimation Accuracy in Orthopantomography. Frontiers in Public Health, 10.
[15] C. Venkatesh, J. Chinnababu, Ajmeera Kiran, C. H. Nagaraju, and Manoj Kumar (2023). A hybrid model for lung cancer prediction using patch processing and deeplearning on CT images. Multimedia Tools and Applications.