IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection Using Machine Learning

Dr.Naresh Thoutam[1], ,Vipul Lokhande[2], Aaditi Baviskar[3], Vidur Diwate[4], Hitesh Pawade[5]

Dept. of Computer Engineering,Sandip Institute of Technology & Research Center,Nashik, India [1],

Assistant Professor, Dept. of Computer Engineering Sandip Institute of Technology & Research Center,Nashik, India

Abstract: This study aims to detect lung cancer at its early stages and evaluate the accuracy of various machine learning models used in this process. Through an extensive review of existing literature, it was observed that some classification algorithms yield lower accuracy, while others perform better but still fall short of achieving near-perfect results. One of the major challenges is the high implementation cost and low accuracy, often due to the improper handling of DICOM images. In medical imaging, several types of scans are available, but CT (Computed Tomography) images are preferred due to their reduced noise levels. Deep learning has emerged as the most effective approach for tasks like lung nodule detection, classification, feature extraction, and predicting the stage of lung cancer. The proposed system initially applies image processing methods to isolate the lung region. Segmentation is carried out using the K-Means clustering algorithm. After segmentation, features are extracted and classified using multiple machine learning techniques. The efficiency of these methods is assessed based on accuracy, sensitivity, specificity, and classification time.

Keywords: Structural Co-occurrence Matrix (SCM), Classifier, Dataset, ROC Curve, Malignant Nodule, Benign Nodule

I. INTRODUCTION

Lung cancer remains difficult to prevent due to its uncertain causes, making early detection essential for effective treatment. The stage of lung cancer is determined by tumor size and the extent of its spread. Globally, lung cancer is a leading cause of mortality, with a 5-year survival rate ranging from just 10% to 16%. Often, lung nodules are not easily visible, requiring expert interpretation and substantial time for diagnosis. Furthermore, many nodules are non-cancerous and can be caused by benign growths, scars, or infections. Although numerous studies have employed machine learning techniques, achieving optimal performance remains difficult due to the need for manual tuning of multiple parameters, making it challenging to reproduce successful results. Classification plays a crucial role in organizing images based on similarity, which becomes complex in cancer cells where overlapping occurs frequently. Therefore, detecting cancer at an early stage remains a significant challenge. After reviewing various studies, it was noted that ensemble classifiers outperform individual machine learning models. Existing computer-aided diagnosis (CAD) systems that rely on CT images for early lung cancer detection have often shown limited success due to low sensitivity and a high false positive rate (FPR).

II. LITERATURE REVIEW

1. In reference [11], Pankaj Nanglia and Sumit Kumar introduced a hybrid technique known as the Kernel Attribute Selected Classifier. This approach combines Support Vector Machine (SVM) with Feed-Forward Back

2. Propagation Neural Network (FFBPNN), aiming to reduce the computational complexity involved in classification. Their model operates in three main stages: the first step is data preprocessing; the second involves feature extraction using the SURF technique, followed by optimization through genetic algorithms; and the final step is classification using FFBPNN. Their method achieved an accuracy of 98.08%.

3. In reference [12], Chao Zhang, Xing Sun, and Kang Dang performed a sensitivity analysis using data from multiple centers. The dataset was categorized based on diameter and pathological results. Diameter was subdivided into three ranges: 0–10mm, 10–20mm, and 20–30mm. For the 0–10mm group, sensitivity was 85.7% (95% CI: 70.8%–100.0%) and specificity was 91.1% (95% CI: 86.8%–95.2%). In the 10–20mm group, the sensitivity was 85.7% (95% CI: 77.1%–94.3%) and specificity was 90.1% (95% CI: 84.8%–95.4%). For the 20–30mm group, sensitivity stood at 78.9% (95% CI: 66.0%–91.8%) and specificity was 91.3% (95% CI: 83.2%– 99.4%). Their model achieved the highest classification accuracy of 85.7% for adenocarcinoma and 65.0% for squamous cell carcinoma.

4. In reference [13], Nidhi S. Nadkarni and Prof. Sangam Borkar centered their research on differentiating between normal and abnormal lung images. They utilized a median filter to eliminate impulse noise, which improved image quality prior to classification. Their approach focused on enhancing diagnostic precision by preprocessing and classifying CT images more effectively.

5. Morphological operations play a crucial role in accurately segmenting the lung region and identifying potential tumor areas. From the segmented images, three geometric features—**area**, **perimeter**, and **eccentricity**—are extracted and used as input for classification through a Support Vector Machine (SVM) model.

6. In reference [14], Ruchita Tekade and Prof. Dr. K. Rajeswari explored lung nodule detection and malignancy prediction using CT scan images. Their study utilized datasets including **LIDC-IDRI**, **LUNA16**, and **Data Science Bowl 2017**, and was executed on a CUDA-enabled GPU (Tesla K20). They employed an **Artificial Neural Network (ANN)** for both feature extraction and classification. Lung nodule segmentation was performed using the **U-NET architecture**, and a 3D multi-graph structure similar to VGG was implemented for classifying the nodules and predicting malignancy levels. The combination of these methods yielded promising results with an **accuracy of 95.66%**, a **loss of 0.09**, and a **Dice coefficient of 90%**, while the **log loss for malignancy prediction** was recorded at **38%**.

7. In reference [15], Moffy Vas and Amita Dessai focused on distinguishing between cancerous and non-cancerous lung images. Their methodology began with preprocessing the CT images to remove irrelevant regions. They applied a **median filter** to eliminate salt-and-pepper noise and utilized **mathematical morphological operations** for precise lung segmentation and tumor detection. Seven features were extracted—**energy, correlation, variance, homogeneity, difference entropy, information measure of correlation**, and **contrast**. These features were fed into a **feed-forward neural network** trained using the **backpropagation algorithm**. The model minimized error using gradient descent with adaptive weight adjustments. Their system achieved a **training accuracy of 96%**, **testing accuracy of 92%**, **sensitivity of 88.7%**, and **specificity of 97.1%**.

8. In reference [16], Radhika P.R. and Rakhi A.S. Nair conducted research on the classification and prediction of medical imaging data. They used datasets from the **UCI Machine Learning Repository** and **Data.World** to evaluate multiple machine learning algorithms. Their findings highlighted that **Support Vector Machine (SVM)** achieved the highest classification accuracy at **99.2%**, followed by **Decision Tree (90%)**, **Naïve Bayes (87.87%)**, and **Logistic Regression (66.7%)**.

9. In reference [17], Vaishnavi D., Arya K.S., Devi Abirami T., and M.N. Kavitha proposed a detection algorithm for lung cancer. During preprocessing, they applied the **Dual-Tree Complex Wavelet Transform (DTCWT)**, a discrete sampling technique for wavelet transformation. For texture analysis, they used the **Gray-Level Co- occurrence Matrix (GLCM)**, a second-order statistical tool that quantifies the frequency of pixel intensity combinations. A **Probabilistic Neural Network (PNN)** was utilized for classification, which demonstrated fast training and high classification accuracy.

10. In reference [18], K. Mohanambal and Y. Nirosha utilized the **Structural Co-occurrence Matrix (SCM)** for feature extraction from CT scan images. Based on these features, the lung nodules were classified as **malignant** or **benign** using an **SVM classifier**, which also categorized the nodules by malignancy level ranging from **1 to 5**.

III. METHODOLOGY

I. SYSTEM MODEL

II. DATA EXPLORATION

This study makes use of **three major datasets**, all containing annotated nodule locations for segmentation tasks, along with labels indicating cancer or non-cancer status for classification purposes.

1. TCIA
2. Dataset

The **Cancer Imaging Archive (TCIA)** provides a collection of anonymized medical imaging data, mostly in **DICOM format**. These collections are categorized based on disease types and imaging modalities, such as CT or MRI. The CT image data supporting this study is sourced from the **Lung CT-Diagnosis repository** (doi.org/10.7937/K9/IA.2015.A6V7JIWX).

3. LIDC-IDRI
4. Dataset

The **Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)** offers a dataset comprising CT scans from **1,018 patients** (total size of 124 GB). The scans were reviewed and annotated independently by four experienced radiologists, highlighting the nodules.

5. **Kaggle Data Science Bowl 2017 Dataset**

This dataset includes CT scans from **1,595 patients** (approximately 146 GB). It provides labels indicating whether each patient was diagnosed with lung cancer in the future, up to a year after the CT scans were initially taken.

A. ALGORITHMS AND TECHNIQUES

For segmenting biomedical images, particularly lung nodules, the **U-Net Convolutional Neural Network** is utilized. This network is designed to process an input image and generate an output mask that identifies the region of interest. Initially, it extracts a set of features through convolutional layers, and subsequently, an up sampling network reconstructs the segmentation mask from these feature vectors [20][21][22]. This represents a **binary classification** problem where **morphological and radiological features** are derived from both the original images and their corresponding masks. While these features are inherently continuous and numerical, they can also be transformed into categorical data when needed. The study explores a range of classifiers for this purpose [23][24][25]:

1. **Logistic Regression** – Well-suited for binary classification tasks, logistic regression often performs effectively in scenarios like this, making it a strong candidate for modeling.
2. **Gaussian Naïve Bayes** – This algorithm works well with continuous data. It computes the mean and variance for each feature within each class, making it appropriate when the feature distribution is close to normal [26].
3. **Random Forest** – A popular ensemble method, especially in competitions such as Kaggle, Random Forest creates multiple decision trees using different subsets of features and data. The final prediction is based on a majority vote, helping to reduce the risk of overfitting.
4. **Multinomial Naïve Bayes** – Ideal for categorical inputs, this method transforms continuous features into discrete bins. It may outperform Gaussian NB when the feature distribution is not normal. For instance, features like nodule diameter in non-cancerous cases often display a left-skewed distribution [27].

5. **Support Vector Machine (SVM)** – SVM aims to find the optimal boundary that best separates classes in a high-dimensional feature space. By applying kernel functions, it can model complex decision boundaries [28].
6. **Gradient Boosting** – Another widely-used ensemble method, Gradient Boosting improves performance by focusing on samples that previous trees misclassified. Unlike Random Forest, which uses random subsets, Gradient Boosting builds trees sequentially, refining predictions at each step.
7. **Ensemble Models** – By combining predictions from multiple individual models, ensemble methods can improve accuracy. The final output is typically derived by averaging the results from several classifiers mentioned above.

B. MODEL EVALUATION AND VALIDATION

Model 1: U-Net Convolutional Neural Network for Nodule Segmentation

This model employs the U-Net architecture to segment lung nodules from CT scan images. The segmentation performance is evaluated by comparing the **processed CT image**, the **actual ground truth label**, and the **predicted segmentation label**.

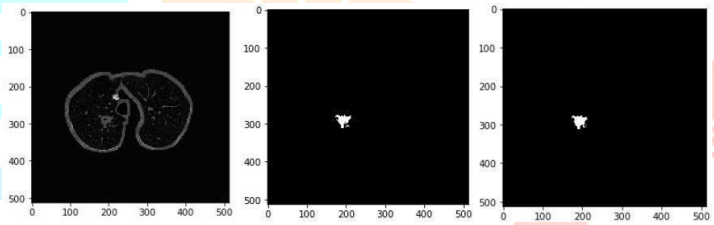


Figure 1: U-Net Image Segmentation – Processed CT scan (left), Ground Truth Label (center), Predicted Label (right)

IV. RESULT DISCUSSION

The dataset was divided using an 80/20 train-validation split via the `train_test_split` function. Due to the extensive training time of approximately 3 hours for just two epochs, cross-validation was not implemented. The **U-Net model** demonstrated convergence after 10 epochs, achieving a **Dice coefficient** of **0.678**, which reflects a 67.8% overlap between predicted nodule masks and their corresponding ground truth labels. Additionally, around **78%** of the predicted masks shared at least one pixel with the actual nodule masks, indicating a decent level of localization accuracy.

The key objective of this study was to accurately pinpoint the **location of lung nodules**, and to assess **sensitivity** and **false positive rates (FPR) per scan** [30][31][32]. Initially, a high number of **false positives (FP)** per true positive (TP) was observed. However, these were significantly reduced in the subsequent model iterations.

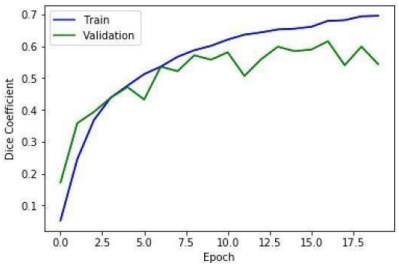
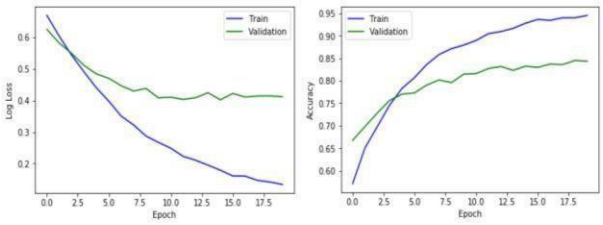


Figure 2 illustrates that a Dice coefficient of **0.678** was achieved, confirming a 67.8% spatial overlap between predicted and actual masks.



In the final model, **hand-selected features** such as **Diameter**, **Spiculation**, **Mean Hounsfield Unit (MeanHU)**, and **Eccentricity** were used for prediction. These were identified through A/B testing as the most impactful features in the highest-performing model.

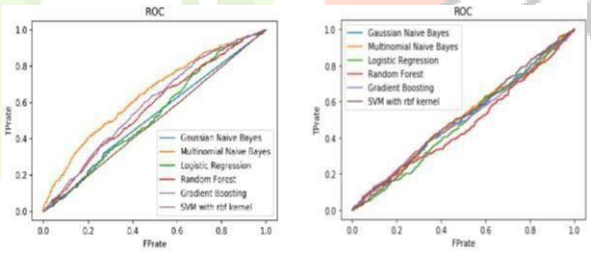
This model outperformed the CNN when evaluated with traditional metrics. It achieved a **log loss** of approximately **0.55**, an **AUC (Area Under Curve)** of **0.64**, and an **average precision** of **0.41**. For comparison, models trained on **random labels** yielded a log loss of **0.59**, AUC of **0.50**, and average precision of **0.29** [33][34].

Considering the **cancer prevalence rate** in the dataset is 26% (or 0.26), the models with randomized labels performed near the baseline expectation. However, models using **true labels** demonstrated substantially **higher performance**, confirming that the selected features hold predictive value

Table 1: Sensitivity, True Positives (TP), and False Positives (FP) Per Scan

A range of classifiers delivered comparable performance once optimized via a **grid search** strategy. This implies a general consistency in the models' capacity to extract relevant information from the input features and generate accurate predictions [35][36]. Moreover, converting continuous features into **discrete bins** (categorical values) by rounding resulted in less than a **0.05% increase in log loss**, underlining the **robustness** of the classification models.

Figure 4:



ROC curves illustrating the **True Positive Rate** versus **False Positive Rate** for models trained with **authentic labels** (left) and **randomized labels** (right).

Model 4: Convolutional Neural Network (CNN) for Cancer Classification Using Detected Nodules

In this model, a CNN was trained to predict whether detected nodules were cancerous or not. The network achieved a **validation loss** of **0.5646** and an **AUC** (Area Under the Curve) of **0.6231**, which is comparable but slightly lower than the performance seen with classifiers trained on hand-selected features. One reason for this could be the reliance on **diameter** as a key feature for cancer detection. While CNNs are highly effective in capturing patterns and textures, they are inherently designed to be **scale and size-invariant**, and may underemphasize spatial dimensions like diameter.

Model	Log Loss True Label	Log Loss Random Label	AUC_ True Label	AUC_ Random Label	Average Precisio n_TI.	Average Precision _RL
Gaussian Naïve Bayes	0.5850	0.8037	0.6380	0.5053	0.4145	0.2929
Multinomial Naïve Bayes	0.5528	0.5920	0.6457	0.5050	0.4100	0.2093
Logistic Regression	0.5525	0.5939	0.6548	0.4823	0.4132	0.2655
Random Forest	0.5533	0.6038	0.6150	0.4681	0.3769	0.2624
Gradient Boosting	0.5672	0.5964	0.6173	0.5019	0.3274	0.2862
SVM-rbf kernel	0.5893	0.5931	0.5017	0.5108	0.2514	0.3787
Ensemble*	0.5519		0.6459		0.4133	

Figure 5:

ROC curves illustrating the **True Positive Rate** versus **False Positive Rate** for models trained with **authentic labels** (left) and **randomized labels** (right).

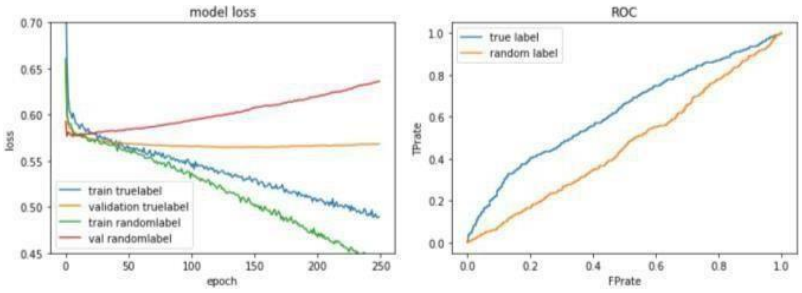
3.1Population and Sample

KSE-100 index is an index of 100 companies selected from 580 companies on the basis of sector leading and market capitalization. It represents almost 80% weight of the total market capitalization of KSE. It reflects different sector company's performance and productivity. It is the performance indicator or benchmark of all listed companies of KSE. So it can be regarded as universe of the study.Non-financial firms listed at KSE-100 Index (74 companies according to the page of KSE visited on 20.5.2015) are treated as universe of the study and the study have selected sample from these companies.

The study comprised of non-financial companies listed at KSE-100 Index and 30 actively traded companies are selected on the bases of market capitalization.And 2015 is taken as base year for KSE-100 index.

3.2 Data and Sources of Data

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.



CONCLUSION

A **Computer-Aided Diagnosis (CAD) system** for lung cancer typically involves several core stages: **image pre-processing**, **nodule detection**, **segmentation**, **feature extraction**, and **classification** into **benign or malignant** categories. Once the nodules are identified and segmented, the next step involves extracting meaningful features using advanced techniques. These features then serve as inputs to classification algorithms designed to predict the likelihood of cancer.

Both **Convolutional Neural Networks (CNNs)** and **traditional machine learning classifiers** were evaluated in this study, with classifiers based on selected features marginally outperforming CNNs. Notably, the **sensitivity** of nodule detection using a **two-stage neural network** reached approximately **65%**, which aligns with the sensitivity range of radiologists (**51%–81.3%**).

However, one limitation of the neural network-based models was a **higher false positive rate**, averaging **6.78 false positives per scan**, whereas radiologists reported significantly fewer false positives (between **0.33–1.39 per scan**). Despite this, when the system focuses exclusively on the **largest detected nodule** for cancer prediction, the **precision** achieved by the machine learning classifiers was considerably higher at **41%**, compared to the typically **1–2% precision** demonstrated by radiologists.

REFERENCES

[1] Saba, T., Sameh, A., Khan, F., Shad, S.A., & Sharif, M. (2019). Lung nodule identification using a combination of handcrafted and deep features. Journal of Medical Systems, 43(12), 332.

[2] Saba, T., Al-Zahrani, S., & Rehman, A. (2012). Development of an expert system to assist offline clinical decision-making. Life Science Journal, 9(4), 2639–2658.

[3] Kim, J., Lee, H., & Yoon, T. (2017). Automated lung cancer diagnosis using deep convolutional neural networks applied to chest CT images. In Proceedings of the 4th International Conference on Biomedical and Bioinformatics Engineering, ACM, Seoul, Korea, 126–132.

[4] Dela Cruz, C.S., Tanoue, L.T., & Matthay, R.A. (2011). Lung cancer: Insights into epidemiology, causes, and prevention. Clinics in Chest Medicine, 32, 605–644.

[5] Chon, A., Balachander, N., & Lu, P. (2017). Leveraging deep convolutional neural networks for the detection of lung cancer.

[6] Khan, S.A., Nazir, M., Khan, M.A., Saba, T., Javed, K., & Rehman, A. (2019). A support vector machine- based system for lung nodule detection in CT images. Microscopy Research and Technique, 82(8), 1256–1266.

[7] Thabsheera, A.A., Thasleema, T.M., & Rajesh, R. (2019). Review of image processing methods for lung cancer diagnosis using CT scans. Springer Singapore, 43, 413–419. https://doi.org/10.1007/978-981-13-2514- 4_34

[8] Saba, T. (2019). A multi-classifier ensemble approach for automated lung nodule classification. Microscopy Research and Technique, 1–9.

[9] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Texture classification using local binary patterns with multi-resolution and rotation invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987.

[10] Ponraj, D.N., Christy, E., & Sharu, M. (2018). Comparative analysis of LBP and LOOP texture features for CT-based lung classification. 4th International Conference on Devices, Circuits and Systems (ICDCS).