Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.15, No.1, 2024

Classification and Diagnosis of Lung Cancer Based Using CNN with VGG-19

Preethi Kolluru Ramanaiah *
Cloud Architect, Lead of AI initiative Program, Ernst & Young LLP, New York, USA
* E-mail: preethiram4@gmail.com / Preethi.kolluru.ramanaiah@ey.com

Abstract

Lung cancer is a major contributor to cancer-related mortality globally, and timely identification is essential for enhancing patient prognosis. Recently, deep learning methods, specifically Convolutional Neural Networks (CNN), have demonstrated encouraging outcomes in image-based medical diagnosis. The paper suggests utilising a CNN-based method to diagnose lung cancer from a healthcare image dataset, with a specific focus on histopathological image data. The proposed CNN approach utilises the natural hierarchical characteristics found in healthcare imagery to autonomously acquire distinctive features that indicate lung cancer. Transfer learning from extensive image datasets and improving models that have been trained are used to tackle the limitations of limited healthcare image datasets successfully. The CNN model utilising the VGG-19 architecture is developed and tested on a comprehensive dataset of lung cancer patients. Following extensive testing and evaluation, the model demonstrates high accuracy as well as precision and recall in diagnosing lung cancer using medical imaging. Interpretability techniques are utilised to get insights into the model's decision-making process, hence increasing its transparency and therapeutic relevance. The proposed CNN-based technology has the potential to help radiologists and clinicians discover and diagnose lung cancer earlier, leading to better patient care and treatment outcomes.

Keywords: Lung Cancer, Histopathological image, Convolutional Neural Networks, VGG-19, Deep Learning
DOI: 10.7176/CEIS/15-1-04
Publication date: March 31st 2024

1. Introduction

Lung cancer is a widespread and fatal type of cancer that presents a major public health issue globally. Timely identification and precise forecasting of lung cancer are essential for enhancing patient results and decreasing rates of mortality. Due to developments in the field of medical imaging and deep learning techniques, there is an increasing interest in creating predictive models for diagnosing and predicting the outcome of lung cancer [1]. Early and precise detection and diagnosis of lung cancer are crucial for enhancing medical results and rates of survival. Medical imaging, namely histopathological pictures, is crucial for identifying and describing lung cancer. Recognizing these images can be difficult, frequently necessitating significant knowledge from radiologists and physicians [2].

Deep learning techniques, such as Convolutional Neural Networks (CNN), have made a significant impact on the field of medical image processing in recent years. Convolutional Neural Networks (CNNs) have demonstrated remarkable ability to interpret complex visual patterns and extract meaningful information from images, making them ideal for image-based medical diagnosis [8]. Researchers are investigating the possibility of hierarchical representation learning in CNNs for lung cancer detection. This study aims to utilise Convolutional Neural Networks (CNNs) for the diagnosis of lung cancer by analysing histopathology images. We want to develop a strong and precise computational approach by merging CNN with VGG-19 to assist in early identification of potential lung lesions and differentiation between benign and malignant nodules. Utilising deep learning algorithms in this scenario has the potential to enhance the abilities of radiologists, enabling more effective and accurate analysis of medical imaging data.

The main contributions of this study on the Classification and Diagnosis of Lung Cancer are outlined below:

- The research focuses on creating and utilising Convolutional Neural Network (CNN) architecture designed specifically for predicting lung cancer using medical imaging data, with a special emphasis on histopathological images.
- The research intends to increase the efficiency and accuracy of lung cancer diagnosis by leveraging CNNs' ability to extract complicated patterns and representations from medical images.

- The incorporation of CNN-based predictive models for lung cancer prediction is an essential step in incorporating deep learning technologies into healthcare.

The literature study encompasses an examination of diverse methodologies employed in the categorization and identification of lung cancer through the utilisation of image processing and classification techniques, as referenced in Section 2. The classification and prediction techniques of the proposed CNN model are presented in the methodology section, as shown in Section 3.

2. Literature Review

In the study of machine learning, Abdullah et al. [8] studied the accuracy ratios of three different classifiers for early-stage lung cancer, with the goal of increasing life-saving efforts. The classifiers under examination include K Nearest Neighbour (KNN), CNN, and SVM. The majority of the relevant indexes used in this investigation were derived from UCI databases, notably those referring to people diagnosed with lung cancer. This study's primary focus is on the execution of these findings, with the ultimate goal of evaluating the performance of classification algorithms utilising the WEKA Tool.

Palani and Venkatalakshmi [3] made a prediction for lung cancer by employing continuous monitoring techniques. The accomplishment was attained by the utilisation of fuzzy cluster-linked classification and enhancement techniques. Fuzzy clustering is necessary for precise image segmentation. The image of lung cancer was analysed using Fuzzy C-means clustering to differentiate the transitional zone. The Otsu thresholding approach was utilised in the study to differentiate between benign and malignant lung tissue. The application of the morphological thinning technique is employed on the right edge image to enhance the optimisation of segmentation presentation. The unique incremental classification technique combines decision tree (DT), and CNN for incremental classification.

In their investigation, Joon et al. [4] used an adaptive spline model to segment lung cancer. The lung X-ray images were produced using X-ray photography using this specific approach. During the preprocessing step, it is recommended to utilise a median filter for noise detection. In the segmentation step, additional techniques such as K-means and fuzzy C-means clustering are used to efficiently capture relevant characteristics. The final result of feature retrieval is obtained after segmenting the X-ray image in the current study. The support vector machine (SVM) approach was used to create the best model for classification purposes.

A comprehensive dataset of labelled lung CT images, encompassing both benign and cancerous instances of lung cancer, was compiled by Vij and Kaswan [5]. The provided photos are utilised as input for the training of Convolutional Neural Network (CNN) models, enabling them to acquire the ability to differentiate between several categories of lung nodules that are suggestive of malignancy. The study examined the structure and configuration of the CNN model utilising VGG-16 for the purpose of predicting lung cancer. The training approach outlined by the authors entails inputting lung CT images into the Convolutional Neural Network (CNN) models and subsequently fine-tuning the model parameters in order to minimise prediction errors. The VGG-16 CNN model attained an accuracy of 77% in this experiment. Sheriff et al. [6] also conducted the implementation of VGG 16, which obtained results that primarily indicated the presence or absence of lung cancer with limited accuracy. Additionally, the study examined various types of lung cancer to assess the severity of its consequences and identify necessary preventive measures.

An investigation into a number of CNN models, such as VGG16, ResNet50V2, and DenseNet201, which are predicated on transfer learning was carried out by M. Phankokkruad et al. [7]. Each model was anticipated to have an accuracy of 63%, 91%, and 91%, respectively, according to the forecasts.

3. Methodology

The approach for the classification of lung cancer using Convolutional Neural Network (CNN) with VGG-19 architecture is primarily built with a few steps, which include the collection of datasets, the preprocessing of images, the training of model, and other stages. Figure 1 has an overview of the structure of the methodology that has been proposed.

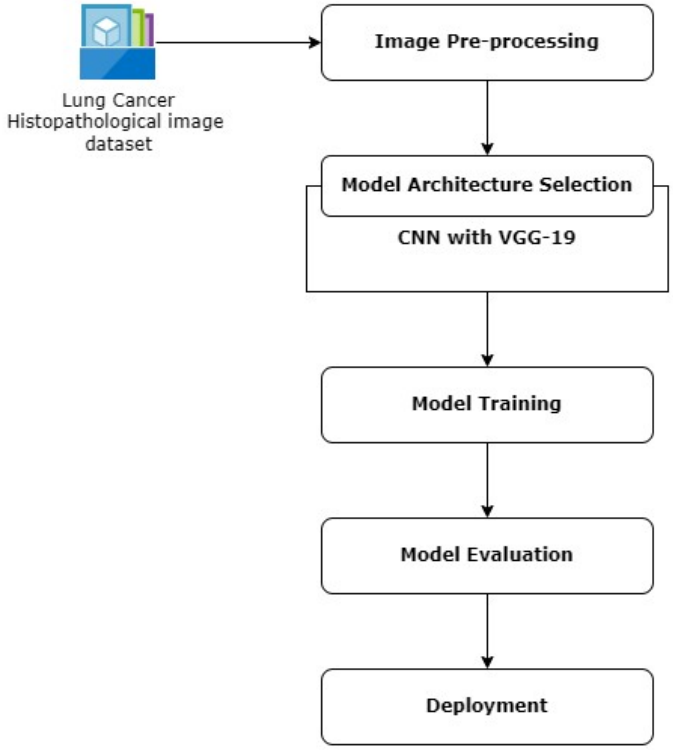
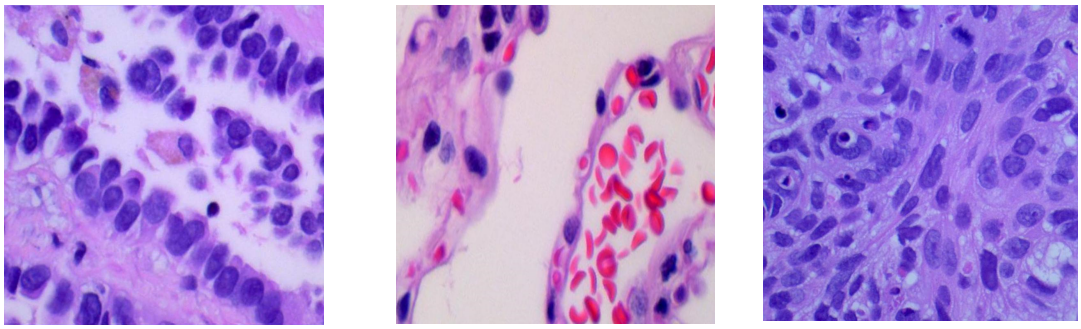


Figure. 1 Architecture of proposed methodology.

3.1 Dataset Collection:

This section includes the collection of a comprehensive dataset comprising lung cancer images, encompassing both malignant and benign instances. In order to ensure effective learning of the CNN model, it is crucial that the dataset is both diverse and well-labelled. Images of lung cancer histopathology were gathered from Kaggle [10] in the data collection phase of this study. The collection boasts a total of three thousand histopathology images of annotated lung cancer. There are images of squamous carcinoma, benign carcinoma, and adenocarcinoma included. All of the students have one thousand JPEG images that are 768 x 768 pixels in size.

The dataset under consideration has been categorised into three distinct classifications, specifically adenocarcinoma, benign, and squamous carcinoma. Figure 2 illustrates the histopathological images of the three classes from the lung cancer dataset. Table 1 described the classes of the lung cancer dataset.



a) Adenocarcinoma b) Benign c) Squamous carcinoma

Figure. 2 Samples of the lung cancer histopathology images

Table. 1 Dataset Description

Class	Description
Adenocarcinoma	Adenocarcinoma is a prominent subtype of non-small cell lung cancer (NSCLC), the predominant form of lung cancer. Adenocarcinoma commonly originates in the peripheral regions of the lungs and exhibits a comparatively slower growth rate in comparison to alternative forms of lung cancer, such as carcinoma of squamous cells.
Benign	Benign refers to non-cancerous conditions that do not infiltrate adjacent tissues or metastasize to other anatomical sites. In contrast to malignant tumours, which possess the ability to become cancerous and spread (metastasize) to adjacent tissues, benign tumours often remain confined to a specific area and do not present a substantial risk to health or life.
Squamous carcinoma	Squamous carcinoma is a primary subtype of non-small cell lung cancer (NSCLC), alongside adenocarcinoma, within the realm of lung cancer. Squamous cell carcinoma commonly originates within the interstitial spaces of the pulmonary system, specifically the bronchi. It is frequently linked to a past of tobacco consumption and contact with cancer-causing substances present in tobacco smoke.

3.2 Image pre-processing:

In order to classify images, algorithms are unable to directly comprehend the images. Consequently, it is essential to transform images into pixel format. For that, this work used the Python module Numpy in order to extract features from the images. In the subsequent stages, the image collection will be partitioned into distinct dependent and independent variables. In this context, independent features refer to picture pixels that are recorded in a list. On the other hand, dependent values, such as illness names, classified values, or target values, are considered dependent variables and can likewise be maintained in a separate list.

3.3 Model Architecture Selection:

VGG-19 is an effective convolutional neural network (CNN) structure renowned for its extensive layers and exceptional efficacy in tasks related to image classification. The architecture comprises several convolutional and pooling layers, which are then followed by fully linked layers. The convolutional neural network (CNN) architecture will be constructed using the VGG-19 model, which consists of 16-layer inputs and 3 layers for output. Subsequently, the VGG-19 model will undergo training using a designated training set, resulting in the generation of a training model that will be employed for subsequent lung cancer diagnosis.

3.4 Model Training:

There are four steps involved in this level. In this article, we will take a quick look at the training method that the CNN uses with VGG-19 [11]:

Input: 224×224 pixels is the size of the image that is taken in by the VGGNet. . This was done to ensure that the input size of the image remained consistent pre trained model of ImageNet competition.

Convolutional Layers: Convolutional layers, max-pooling layers, and fully connected layers are some of the 19 layers that make up VGG-19, which is just what its name says. When it comes to acquiring characteristics based on the input images, the convolutional layers are the fundamental building blocks that are essential. The neural network is structured into blocks, with each block including multiple convolutional layers, which are then followed by max-pooling layers. Tiny 3x3 filters with a stride of 1 and no padding are generally utilised by the convolutional layers in order to achieve the goal of preserving the spatial dimensions of the inputs. As we progress farther into the network, the quantity of layers that are contained within each convolutional layer grows. This enables the model to acquire knowledge about features that are ever more complicated. There are a total of sixteen convolutional layers in VGG-19, which are arranged into five ConvBlocks.

Max-Pooling layers: Following each group of convolutional layers, the VGG-19 algorithm incorporates max-pooling layers that have a window size of 2x2 and a stride of 2. The spatial dimensions of the feature maps are down sampled using max-pooling layers, which helps to reduce the complexity of the computations involved and also offers assistance in capturing translation-invariant characteristics.

Fully-connected layers: The layers in question are comprised of neurons that are highly interconnected, meaning that each neuron within a given layer is linked to all of the neurons in the layer before it. As the classification process progresses, the total amount of neurons in the fully linked layers exhibits a steady reduction towards the output layer, aligning with the amount of output classes or categories.

3.5 Model Evaluation:

The assessments of performance will be prepared with the assistance of evaluating the dataset of images during the process. In this case, the trained CNN model will be utilised to compute outcomes such as accuracy, loss, precision, and recall by utilising the testing dataset as the input.

3.6 Deployment:

When the model has reached a level of performance that is satisfactory, it will be able to be used for the classification of lung cancer images. The detection of lung cancer is accomplished through the use of a histopathology image as an input file during the deployment stage. This system extracts the distinctive qualities of the image from the input image, and then it feeds these features into the CNN, a deep learning model. The proposed CNN model is used for the classification of lung cancer images in the deployment stage.

4. Results

This study collected malignant and benign lung cancer photos for the experiment. The dataset must be diverse and well-labeled for CNN model learning. Kaggle [10] was used to acquire lung cancer histology images for this investigation. The collection includes 3,000 annotated lung cancer histology photos. In order to facilitate the training and testing processes, the dataset will be divided in a ratio of 70 to 30. The training set is represented by 70 percent of the total, while the testing set is represented by 30 percent. The CNN model is evaluated by using the testing dataset as the input in order to compute outcomes such as accuracy, loss, precision, and recall. This procedure is carried out in order to evaluate the CNN model.

The results analysis section presents a comparative evaluation of the lung cancer prediction model's performance in two sets of experiments, each with varying epoch values. The analysis focuses on the convergence behaviour of the model in relation to the number of training epochs. Convergence is the point at which the training process reaches a state of stability, where the performance of the model reaches a plateau and subsequent training repetitions result in declining improvements. Insights into the appropriate training length for achieving satisfactory performance can be obtained by comparing the convergence patterns found with 20 and 30 epochs. The performance scores with epoch 20 were reported in Table 2, and the results of performance scores with epoch 30 were reported in Table 3. In the same way, the result graph of the performance scores with epoch 20 is illustrated in Figure 3, and the results of the performance scores with epoch 30 is illustrated in Figure 4.

Table. 2 Performance results of the Proposed CNN with VGG-19 (epoch 20)

Accuracy	**Loss**	**Precision**	**Recall**
0.93	0.17	0.93	0.93

Table. 3 Performance results of the Proposed CNN with VGG-19 (epoch 30)

Accuracy	**Loss**	**Precision**	**Recall**
0.94	0.13	0.94	0.94

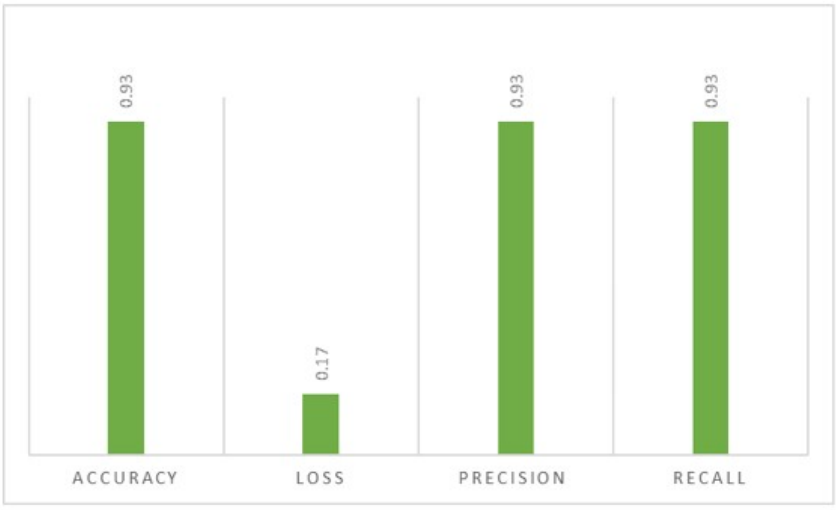


Figure. 3 Graph representation of Performance results of the CNN with VGG-19 (epoch 20)

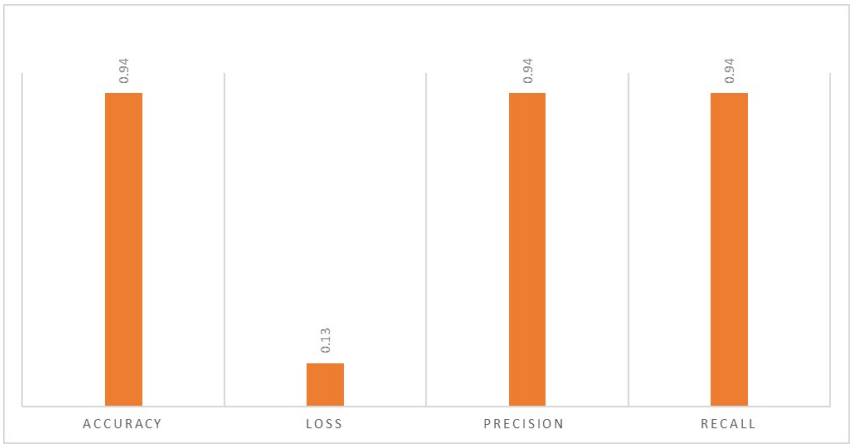


Figure. 4 Graph representation of Performance results of the CNN with VGG-19 (epoch 30)

Therefore, the results chapter consolidates the outcomes of the experiments conducted on two different time periods (20 and 30) and clarifies their significance in enhancing the accuracy of the lung cancer prediction model, outperforming earlier studies.

Conclusions

When compared to various algorithms for machine learning, deep learning has a number of advantages, one of the most significant of which is its capacity to carry out feature engineering tasks on its own. In order to allow faster learning, this evaluates the data in order to uncover similar qualities and then integrates those features. In recent years, deep learning techniques like CNNs have transformed medical image processing. Convolutional Neural Networks (CNNs) are perfect for image-based medical diagnosis because they can analyse complicated visual patterns and extract useful information from images. The CNN algorithm will identify the input lung picture as either normal or abnormal when it has successfully completed the training and testing phases. For this reason, a Deep Convolutional Neural Network (CNN) with the VGG-19 model was utilised for this proposed methodology to predict lung cancer detection using histopathological images. The VGG-19 is not a distinct

model from CNNs; rather, it is a particular kind of architecture that is exclusive to CNNs. When comparing CNN architectures, it is highly important to take into consideration how well they function. On the basis of the findings of this experiment, it has been demonstrated that the CNN with VGG-19 has obtained superior performance in comparison to the CNN results. It is suggested that future research involve the incorporation of various additional image datasets pertaining to lung cancer and the identification of a more precise model to predict lung cancer based on the images.

References

[1] Hatuwal, Bijaya & Thapa, Himal. (2020). Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. *International Journal of Computer Trends and Technology*. 68. 21-24. 10.14445/22312803/IJCTT-V68I10P104.

[2] Ardila, Diego & Kiraly, Atilla & Bharadwaj, Sujeeth & Choi, Bokyung & Reicher, Joshua & Peng, Lily & Tse, Daniel & Etemadi, Mozziyar & Ye, Wenxing & Corrado, Greg & Naidich, David & Shetty, Shravya. (2019). *End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography*. Nature Medicine. 25. 1. 10.1038/s41591-019-0447-x.

[3] Palani, D. & Venkatalakshmi, K.. (2018). An IoT Based Predictive Modelling for Predicting Lung Cancer Using Fuzzy Cluster Based Segmentation and Classification. *Journal of Medical Systems*. 43. 10.1007/s10916-018-1139-7.

[4] Joon, P., Bajaj, S.B., Jatain, A. (2019). Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Techniques. In: Pati, B., Panigrahi, C., Misra, S., Pujari, A., Bakshi, S. (eds) *Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing*, vol 713. Springer, Singapore.

[5] A. Vij and K. S. Kaswan, (2023). Prediction of Lung Cancer using Convolution Neural Networks, *International Conference on Artificial Intelligence and Smart Communication (AISC),* Greater Noida, India, 2023, pp. 737-741, doi: 10.1109/AISC56616.2023.10085058.

[6]] S. T. M. Sheriff, J. V. Kumar, S. Vigneshwaran, A. Jones, and J. Anand. (2021). Lung cancer detection using vgg net 16 architecture, in *Journal of Physics: Conference Series*, vol. 2040, p. 012001, IOP Publishing.

[7] Manop Phankokkruad. (2021). Ensemble Transfer Learning for Lung Cancer Detection. In 2021 4th *International Conference on Data Science and Information Technology (DSIT 2021).* Association for Computing Machinery, New York, NY, USA, 438–442.

[8] bramanian, R. Raja, R. Nikhil Mourya, V. Prudhvi Teja Reddy, B. Narendra Reddy, and Srikar Amara. (2020). *Lung Cancer Prediction Using Deep Learning Framework. International Journal of Control and Automation*.

[9] Abdullah, Dakhaz & Mohsin Abdulazeez, Adnan & Sallow, Amira. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. *Qubahan Academic Journal*. 1. 141-149. 10.48161/qaj.v1n2a58.

[10] Srinivas, Lung_cancer_preprocessed_Dataset. (2023). Kaggle. [Online] *https://www.kaggle.com/datasets/srinivasbece/lung-cancer-preprocessed-dataset.*

[11] Team, K. (n.d.). Keras documentation: VGG16 and VGG19. [Online] *https://keras.io/api/applications/vgg*