International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Forecasting Lung Cancer Using Convolution Neural Network

Rahul Metuku*[1], Ch. Niranjan Kumar[2], V. Sowmya Devi[3]

Submitted:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract: Lung cancer is a major cause of mortality globally, underscoring the need for precise prediction methods to reduce death rates. The application of artificial intelligence (AI) to CT scan images has shown significant potential in enhancing lung cancer predictions through automated processes. Among AI techniques, deep learning, particularly Convolutional Neural Networks (CNNs), stands out in predictive analytics, frequently outperforming other machine learning algorithms. This study analysed a dataset of 1,000 chest CT scan images representing different types of lung cancer, including Adenocarcinoma, Benign, and Squamous Cell Carcinoma. Various machine learning algorithms were evaluated, with CNNs achieving the highest prediction accuracy. The existing system, which employs the VGG-16 model, achieves an accuracy of 77.62%, which is considered suboptimal. To address this, the proposed system uses the VGG-19 transfer learning model on the dataset, aiming to improve prediction accuracy and offer better insights into the severity and necessary precautions for different lung cancer types.

Keywords: *Lung Cancer, Deep Learning, CNNs, Transfer Learning, Medical Imaging, Predictive Analytics, Diagnostic Accuracy*

1. Introduction

The uncontrolled growth of cells in the lungs is known as lung cancer, a leading cause of mortality globally. It has surpassed the combined fatalities of breast, prostate, and colon cancers. Smoking addiction is a primary cause, but exposure to carcinogenic environments like radioactive gas and air pollution, as well as genetic factors, significantly contribute to lung cancer development. The uncontrolled growth of tissue leads to this disease, with primary lung cancer originating from lung cells and secondary lung cancer starting elsewhere in the body and spreading to the lungs. Lung cancer cells differ markedly from normal cells; they grow rapidly, appear abnormal, and are prone to spreading [1].

Lung cancer is one of the deadliest diseases, with increasing cases, particularly in India, where around 70,000 new cases are reported annually. Its asymptomatic nature in early stages makes early detection crucial for improving survival rates. Early detection offers patients better chances of recovery and cure. Technological advancements play a vital role in early cancer detection. Researchers have developed various computer-aided diagnosis (CAD) systems utilizing machine learning and deep learning techniques [2]. These systems analyse images to predict cancer malignancy levels. This study aims to leverage transfer learning techniques to analyse image datasets for the classification and early detection of lung cancer.

[1] *Sreenidhi Institute of Science and Technology, Hyberabad, India*
ORCID ID :
[2] *Sreenidhi Institute of Science and Technology, Hyberabad, India*
ORCID ID : 0000-0001-6827-5770
[3] *Sreenidhi Institute of Science and Technology, Hyberabad, India*
ORCID ID : 0000-0002-5055-2466
** Corresponding Author Email: rahulmetuku@gmail.com*

2. Literature Survey

Several researchers have conducted studies with various outcomes. C. Yao et al. developed a CNN model with a custom dataset, which reached 90% accuracy. However, the reliability of the data and the model's effectiveness on real-world datasets are still in question. This method requires further testing on multiple datasets to confirm its reliability. Similarly, [3] proposed that nanotechnology could be more effective and less toxic for lung cancer therapy compared to traditional chemotherapy. This survey indicated that nanotechnology holds promise, although its effectiveness depends on patient selection and the combination of multiple treatments.

[4] reviewed various studies and observed that high false-positive rates often decrease accuracy. To tackle this problem, L. Ye et al. suggested moving from 2-dimensional to 3-dimensional architectures, such as 3D-CNN, VGG-16, AlexNet, and Multi-Crop Net, achieving an improvement of 8.28%, though this is still lower compared to other algorithms.

A dataset of over 1000 images was analysed [5][6], who focused on Stages T1a-3N0M0 of non-small cell lung cancer (NSCLC) were analysed, revealing that NSCLC tends to be more fatal than small cell lung cancer, with a higher mortality rate. A. Agaimy et al. examined survival rates of patients following an initial NSCLC diagnosis. The study involved 14 patients (8 males and 6 females) aged 52 to 85 years (median age 60). The longest survival recorded was 45 months for a patient with cerebral metastasis.

S. T. M. Sheriff et al. used the VGG-16 model to detect lung cancer, although the accuracy was limited. Building on this research, further studies are focused on not only

detecting the presence of cancer but also identifying its type to assess severity and recommend appropriate precautions [7][8].

[9][10] evaluated several CNN models, including VGG-16, ResNet50V2 and DenseNet201, both utilizing transfer learning. These models achieved accuracies of 62%, 90%, and 89%, respectively.

Advancements in medical imaging technologies, particularly through the use of Convolutional Neural Networks, have enabled more accurate and timely identification of lung cancer symptoms, thereby supporting healthcare professionals in diagnosing this life-threatening disease at earlier stages.[11]

By 2040, 28.4 million people will have hereditary lung cancer, which has unknown causes. For attribute selection, previous research used information gain models, multilayer perceptron, random subspace, and sequential minimal optimisation. Using large parameters and single threshold values can be inefficient. A new lung cancer prediction method uses Z-score normalisation, levy flight cuckoo search optimisation, and weighted convolutional neural network [12].

The paper suggests a two-step verification design that uses a Decision Tree classification and a VGG16 CNN model to find lung cancer. The model looks at a person's symptoms and medical history to figure out how likely it is that they will get lung cancer and then checks this with a CT scan picture [13].

3. Methodology

VGG-19 is an advanced deep convolutional neural network (CNN) with 19 layers that has demonstrated superior performance in image classification tasks. It is structured to process input images through a sequence of convolutional layers, followed by fully connected layers, is used to classify images into different categories. Here is an elaboration on its core components and architecture.

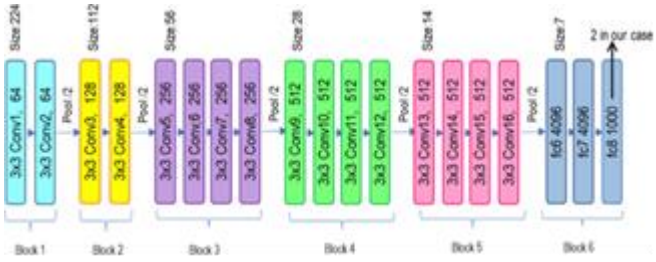

Fig. 1. Architecture of VGG – 19

Input Layer: VGG-19 accepts images of size 224×224 pixels. Images are resized to this dimension for consistency, often by cropping a central patch of the required size.

Convolutional Layers: The convolutional layers in VGG-19 use small receptive fields of 3×3 pixels, which capture fine-grained details in images. These layers apply convolution operations with a fixed stride of 1 pixel, ensuring that the spatial resolution is preserved. The network also includes 1×1 convolution filters, which perform linear transformations on the input data.

Activation FunctionEach convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function. The ReLU function adds non-linearity to the model, aiding in the learning of complex patterns. It directly outputs the input if it is positive; otherwise, it outputs zero.

Pooling Layers: A max-pooling layer is applied following a group of convolutional layers. Max-pooling reduces the spatial dimensions of the input by selecting the maximum value within each region from a cluster of neurons at the previous layer, typically using a 2×2 pooling window. This process reduces computational load and controls overfitting.

Hidden Layers: All hidden layers in VGG-19 use ReLU activations. The network typically avoids Local Response Normalization (LRN) to conserve memory and reduce training time since LRN does not significantly boost accuracy.

Fully-Connected Layers: The network includes three fully connected (FC) layers. The first two layers each have 4096 units, while the third FC layer has 1000 units, corresponding to the number of classes in the ImageNet dataset. The final layer employs a SoftMax activation function to generate class probabilities.

3.1. Training and Transfer Learning

Pre-training on ImageNet: VGG-19 is pre-trained on the ImageNet database, which includes millions of images across 1000 categories. This pre-training enables the model to learn broad features from a large and varied dataset.

Transfer Learning: For specific tasks, such as lung cancer classification, the pre-trained VGG-19 model can be fine-tuned. Transfer learning involves using the pre-trained weights and adapting the final layers of the network to the new task. This process helps in leveraging learned features, speeding up training, and often achieving better performance with less data.

3.2. Evaluation Metrics

Accuracy: Represents the percentage of correctly classified instances relative to the total number of instances.

Loss: Measures the discrepancy between predicted outputs and actual outputs, guiding the optimization process.

Precision: Shows the ratio of true positive predictions to the total number of positive predictions made.

Recall: Indicates the ratio of true positive predictions to the total number of actual positive instances.

3.3. Architectural Details of VGG-19

Blocks: VGG-19 is organized into 5 blocks of convolutional layers, each paired with a max-pooling layer. As the network progresses, the number of filters in the convolutional layers doubles, starting from 64 filters in the first block and increasing up to 512 filters in the fifth block.

Dense Layers: The last three layers are dense (fully connected) layers. The dimensions of these layers are 4096, 4096, and 1000, respectively. For the specific task of lung cancer classification, the final layer dimension can be adjusted to match the number of classes, such as two for binary classification (e.g., cancerous vs. non-cancerous).

3.4. Summary of VGG-19 Benefits

small Filters: Use of small 3×3 filters allows capturing intricate details while maintaining computational efficiency.

Deep Architecture: The depth of VGG-19 enables it to learn complex features from data.

Transfer Learning: Pre-training on a large dataset like ImageNet allows VGG-19 to be effectively adapted to new tasks with relatively smaller datasets.

High Accuracy: The network has demonstrated high accuracy in image classification, making it suitable for medical image analysis tasks like lung cancer detection.

By leveraging the architecture and benefits of VGG-19, this methodology aims to provide a robust framework for accurately classifying lung cancer images and improving early detection and diagnosis.

4. Implementation

4.1. Dataset Collection

Source and Composition:

Acquire the Lung Cancer Histopathological Images dataset from the Kaggle web repository [10][14]. The dataset includes 3000 images divided into three classes: Adenocarcinoma, Benign, and Squamous Cell Carcinoma, with 1000 images per class. Each image is 768 x 768 pixels in size and saved in JPEG format.

Steps:

1. Download the dataset from Kaggle.

2. Organize the images into separate folders for each class.

4.2. Image Pre-processing

Convert raw images into a format suitable for analysis and model training.

Steps:

1. Read Images: Use Python libraries to navigate through the dataset directories and load the images.

2. Convert Images to Numerical Format: Change the images into a format of pixel values that can be processed by the machine learning model

3. Normalize Images: Adjust the pixel values to a range of [0, 1] to standardize the input data

4. Encode Labels: Convert the categorical class labels into numerical values.

4.3. Split Dataset

Divide the dataset into training and testing sets to assess the model's performance on new data

Steps:

Split Data: Allocate 70% of the dataset for training and 30% for testing. The training set is used to develop the model, while the testing set is used to evaluate its effectiveness.

4.4. Training the Model

Train a CNN based on the VGG-19 architecture to classify lung cancer images.

Steps:

1. Load Pre-trained VGG-19: Utilize a VGG-19 model that has been pre-trained on the ImageNet dataset, excluding its top layers.

2. Add Custom Layers: Add fully connected layers to adapt the model for the lung cancer classification task.

3. Freeze Base Model Layers: Retain the weights of the pre-trained VGG-19 layers to leverage learned features.

4. Compile the Model: Configure the model with an optimizer, loss function, and evaluation metrics suitable for classification.

5. Train the adapted VGG-19 model using the training dataset, adjusting parameters such as epochs and batch size to optimize performance.

4.5. Performance Evaluations

Evaluate the trained model's performance using various metrics and visualize its training progress.

Steps:

1. Evaluate Model: Test the trained model on the testing set to obtain accuracy and loss values.

2. Calculate Additional Metrics: Compute precision, recall, and F1-score to provide a detailed performance analysis.

3. Visualize Performance: Plot graphs of accuracy and loss over epochs to visualize the model's training and validation performance.

4.6. Lung Cancer Detection

Implement the trained model for real-time lung cancer detection using new pathology images.

Steps:

1. Upload and Pre-process New Image: Provide an interface for uploading a new pathology image, pre-process it to match the input requirements of the model.

2. Predict Class: Use the trained model to predict the class of the uploaded image and display the result, indicating whether the image is classified as Adenocarcinoma, Benign, or Squamous Cell Carcinoma.

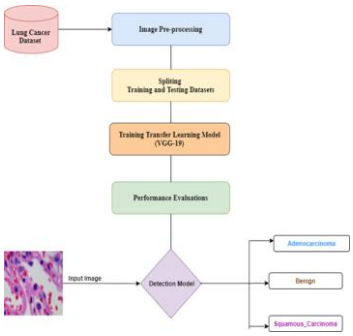

Fig. 2. Data flow diagram

By following these implementation steps, the system can effectively process the dataset, train a robust model, evaluate its performance, and deploy it for real-time lung cancer detection.

5. Results and Observations

5.1. Types of Graphics

The figure illustrated the dataset used in this project is sourced from the Kaggle web repository [14][15], which is renowned for providing datasets for data science and machine learning applications. This particular dataset consists of 3000 histopathological images categorized into three distinct classes: Adenocarcinoma, Benign, and Squamous Cell Carcinoma, with each class containing 1000 images. Each image in the dataset is of high resolution, measuring 768 x 768 pixels, and is stored in JPEG format. This comprehensive collection provides a robust foundation for training and evaluating machine learning models aimed at lung cancer classification and detection.

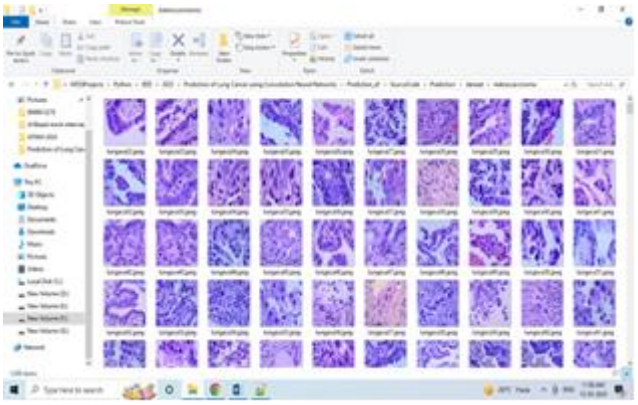

Fig. 3. Source data

The data-set when applied VGG-19 algorithm to check the accuracy and other metrics, resulted in the provided outcomes. Fig. 4. Shows accuracy, Fig. 5. Loss, Fig. 6. Precision and Fig. 7. Recall.

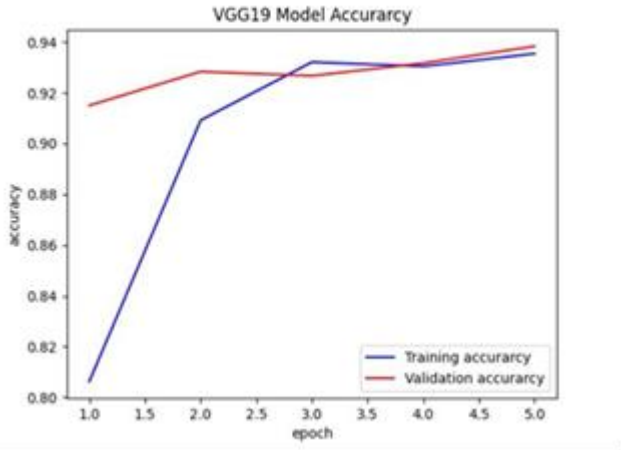

Fig. 4. Training Vs Validation accuracy

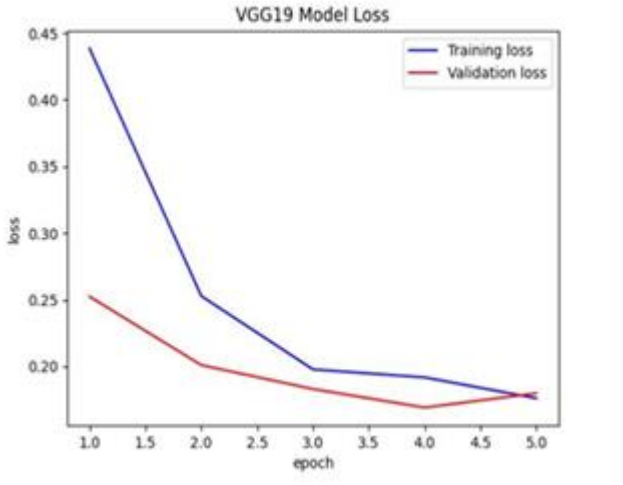

Fig. 5. Model Loss

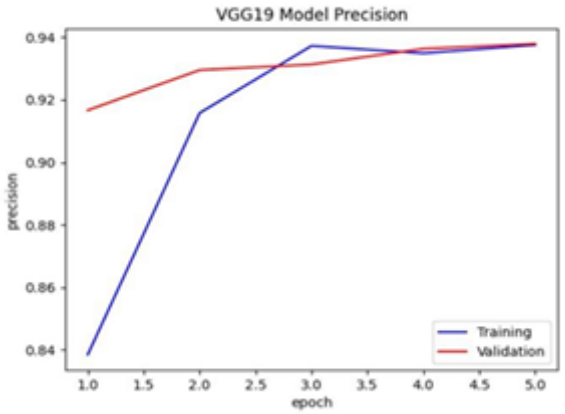

Fig. 6. Model Precision

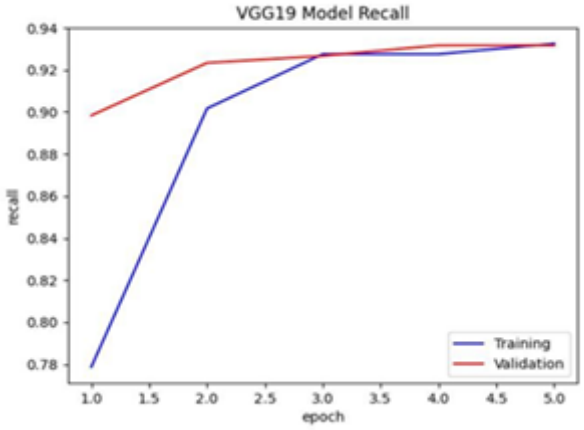

Fig. 7. Model Recall

6. Conclusion

The application of the VGG-19 algorithm to the given dataset has yielded promising results. The model demonstrated strong predictive performance, as evidenced by the high accuracy, recall, precision and less loss. The uphill and downward trends observed in the metrics across the training epochs suggest the model is effectively learning and generalizing from the data. These findings indicate the VGG-19 architecture is well-suited for this particular classification task.

To further enhance the model's capabilities, future work could explore data augmentation techniques, fine-tuning of hyperparameters, and the integration of ensemble methods. Additionally, investigating the transferability of the learned features to related domains could expand the model's applicability. Continuous monitoring and updating of the model as new data becomes available will also be crucial for maintaining its effectiveness over time.

Acknowledgements

We thank our colleague from Sreenidhi Institute of Science and Technology who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

Author contributions

Rahul Metuku: Conceptualization, Methodology **Niranjan Kumar Ch:** Algorithm implementation and simulation **Sowmya Devi V:** Paper drafting, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] M. Norouzi and P. Hardy, “Clinical applications of nanomedicines in lung cancer treatment,” Acta Biomaterialia,vol. 121, pp. 134–142, 2021.

[2] P. H. Viale, “The american cancer society’s facts & figures:2020 edition,” Journal of the Advanced Practitioner inOncology, vol. 11,no. 2, p. 135, 2020.

[3] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M.Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, etal., “Gene-expression profiles predict survival of patients with lung adeno-carcinoma,” Nature medicine, vol. 8, no. 8, pp.816–824, 2002.

[4] C.-R. Guo, Y. Mao, F. Jiang, C.-X. Juan, G.-P. Zhou, and N. Li,“Com- putational detection of a genome instability-derived lncrna signature for predicting the clinical outcome of lung adenocarcinoma,”Cancer Medicine, vol. 11, no. 3, pp. 864–879, 2022.

[5] M. A. Gillette, S. Satpathy, S. Cao, S. M. Dhanasekaran, S. V.Vasaikar,K. Krug, F. Petralia, Y. Li, W.-W. Liang, B. Reva, et al., “Proteogenomic Characterization reveals therapeutic vulnerabilities in lungadenocarci- noma,” Cell, vol. 182, no. 1, pp. 200–225, 2020.

[6] A. Agaimy, O. Daum, M. Michal, M. W. Schmidt, R. Stoehr, A. Hartmann,and G. Y. Lauwers, “Undifferentiated large cell/rhabdoid carcinoma presenting in the intestines of patients with concurrent or recent non-small cell lung cancer (nsclc): clinicopathologic and molecular analysis of 14 cases indicates an unusual pattern of dedifferentiated metastases,” Virchows Archiv, pp. 1–11, 2021.

[7] K. I. Tosios, V. Papanikolaou, D. Vlachos Dimitropoulos, and N.Goutas, “Primary large cell neuroendocrine carcinoma of the parotid gland. report of a rare case,” Head and Neck Pathology, pp. 1–8,2021.

[8] B.-Y. Wang, J.-Y. Huang, H.-C. Chen, C.-H. Lin, S.-H. Lin, W.-H.Hung, and Y.-F. Cheng, “The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients,” Journal of

cancer research and clinical oncology, vol. 146, no. 1, pp. 43–52,2020.

[9] S. Li, P. Xu, B. Li, L. Chen, Z. Zhou, H. Hao, Y. Duan, M. Folkert, J.Ma,S. Huang, et al., "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features," PhysicsinMedicine & Biology, vol. 64, no. 17, p. 175012, 2019.

[10] images from biorender, "www.biorender.com," 2022.

[11] Deep Learning Approach for Early Stage Lung Cancer Detection, Saleh Abunajm, Nelly Elsayed, Zag ElSayed, Murat Özer *arXiv (Cornell University)*, 2023

[12] Koti, Manjula Sanjay, et al. "Lung cancer diagnosis based on weighted convolutional neural network using gene data expression." *Scientific Reports* 14.1 (2024): 3656.

[13] Krishna, S. Udit, et al. "Lung Cancer Prediction and Classification Using Decision Tree and VGG16 Convolutional Neural Networks." *The Open Biomedical Engineering Journal* 18.1 (2024).

[14] Lung and Colon Cancer Histopathological Images (kaggle.com)

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