Project on

Lung Cancer Prediction Using Deep Learning

Masoumeh Khalilzadeh

Abstract

This paper presents a deep learning approach for lung cancer prediction using histopathological images. We design and evaluate transfer learning with the EfficientNetB7 convolutional neural network architecture pre-trained on ImageNet. The histopathological images are preprocessed to enhance their quality and standardized format. We propose a model architecture that integrates global average pooling for feature extraction and fully connected layers for classification. Through extensive experimentation on a dataset of lung images, our approach demonstrates strong predictive performance in accurately identifying lung cancer presence. Additionally, we explore the performance of our model's predictions. Our research contributes to the field by offering a robust and interpretable framework for lung cancer prediction.

Introduction

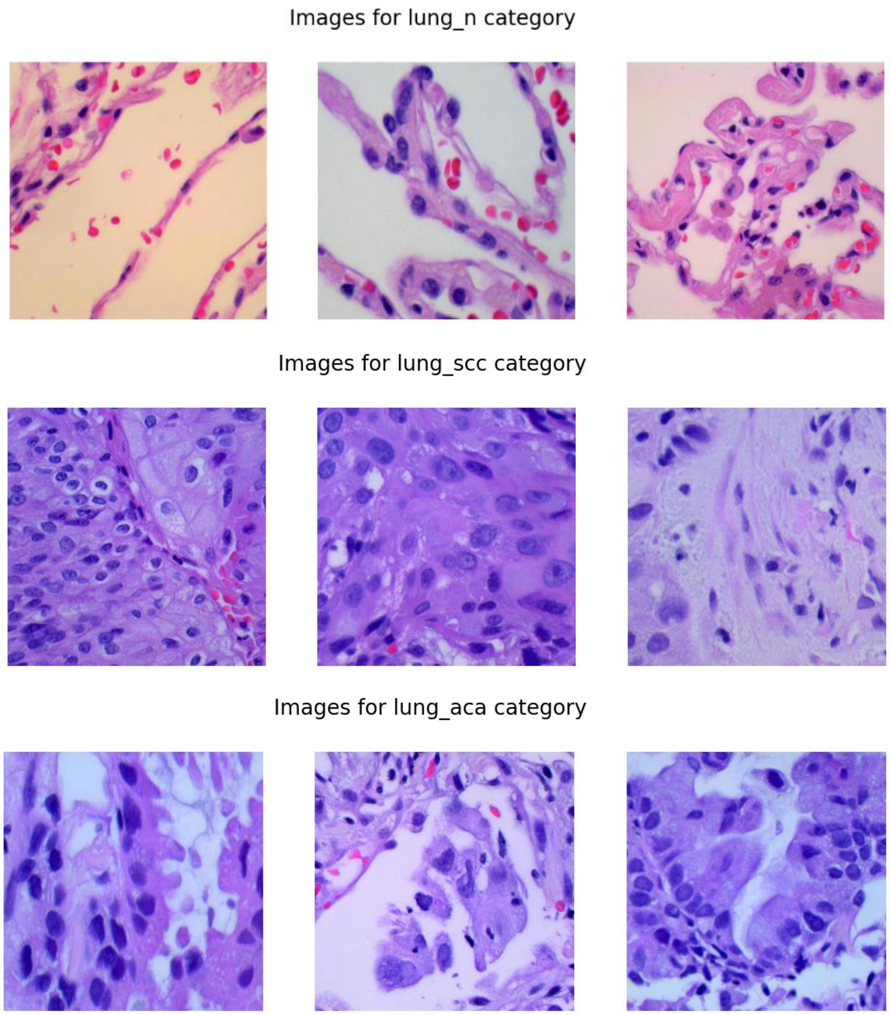
The American Cancer Society estimated that 340 people die each day from lung cancer.[1] Approximately 101,300 of the 125,070 lung cancer deaths (81%) in 2024 will be caused by cigarette smoking directly, with an additional 3500 caused by second-hand smoke.[2] The remaining balance of approximately 20,300 lung cancer fatalities which is not related to smoking would rank as the eighth most common cause of cancer death among both genders combined.[1] The lung cancer rate has shown a consistent decrease since 2006, with a yearly decline of 2.5% among men and 1% among women.[1] There was a notable increase in the death rate from cancer during most of the 20th century, due to a rise in lung cancer especially among men because of using tobacco. [1] Meanwhile, the rate of decline in lung cancer mortality increased from 2% annually between 2005 and 2013 to 4% annually from 2013 to 2021. The Progress in early detection and treatment has a great impact of this rise. [3] So, we can see that the importance of early detection of lung cancer has been increasing, so we should find the best possible ways to have progress in this area. Medical imaging plays an important role in diagnosing various diseases. Image information helps to decide at many stages in the patient care process. For example, in detection, characterization, staging, monitoring of disease recurrence. In recent years, it has been recognized the importance of machine learning and computational techniques in developing computerized methods in image analysis.[4]

Computer-aided diagnosis (CAD) and computer-assisted image analysis have been used as an important tool in performing research and development recently. We can use CAD methods in different areas like disease detection, characterization, staging, treatment response assessment, prognosis prediction, and risk assessment for various diseases and with various imaging modalities. [4] Meanwhile, Deep learning is a form of representation learning technique wherein a multi-layer neural network transforms the input data into multiple layers of abstractions.[5] Furthermore, we can say that CAD application using deep learning has been very useful is areas like classification of disease and normal patterns, classification of malignant and benign lesions, and prediction of high risk and low risk patterns of developing cancer in the future. [4] In this study, our aim is to develop an efficient model for lung cancer prediction by using deep learning algorithm. So, we use convolutional neural networks (CNNs) to identify patterns of lung cancer from histopathological images. Also, our goal is to assess the impact of preprocessing techniques and model evaluation methods on the overall performance of the predictive model.

Using deep learning algorithms for lung cancer detection has been increasing. Recently, researchers have been performing the convolutional neural network in different aspects including computer vision. There are different CNN-based methods for investigating natural image processing and medical image analysis. For example, in [9], a three-dimensional convolutional neural network (CNN) consisting of three modules was developed to identify and classify lung nodules. In [10] an artificial neural network (ANN) has been performed to detect lung cancer. Also, you can find a proposed deep learning method with Trained Neural Networks (DITNN)in [11] they used clustering technique (IPCT), which significantly enhanced lung image quality and achieved lung cancer diagnosis accuracy. Furthermore, [12] presented a double convolutional deep neural network (CDNN) and a conventional CDNN in lung nodule recognition and classification. Two studies ([13, 14]) utilized deep learning algorithms to predict the survival rate of lung adenocarcinoma and determine EGFR mutation status and its subtypes.

Methodology

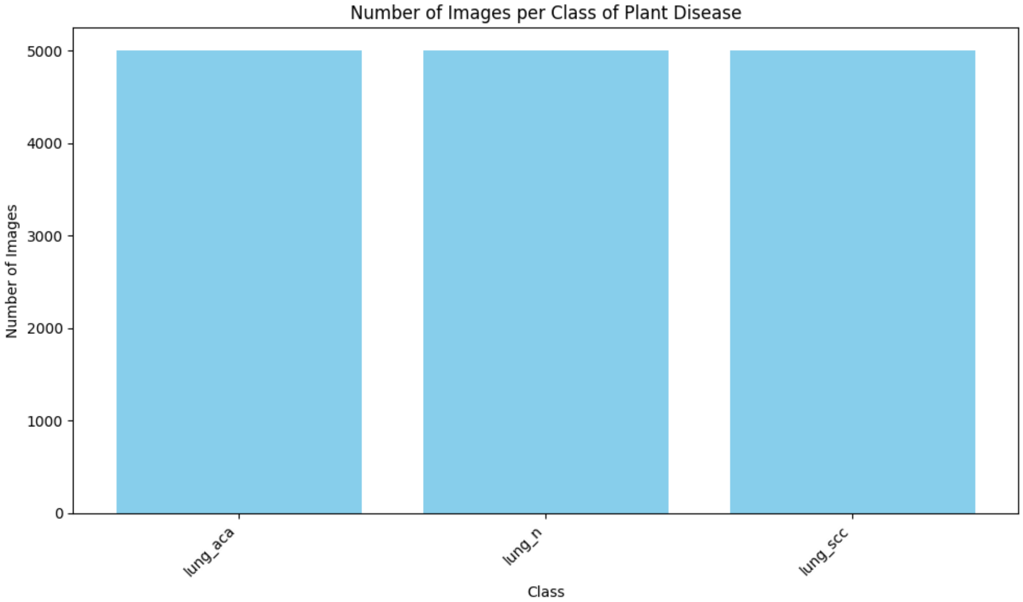
In this research, our dataset contains 15000 histopathological images which is obtained from Kaggle, a public data repository for datasets. This dataset is from a repository titled "Lung and Colon Cancer Histopathological Images" by Andrew Maas. These 15000 histopathological images are presented in three classes based on different lung conditions. These classes are normal lung tissue samples, lung tissue samples with adenocarcinomas which is a type of lung cancer and lung tissue samples with squamous cell carcinomas which is another type of lung cancer. Each of these three classes contains 5000 images and they have been developed by data augmentation from 250 images. So, there is no need to perform data augmentation on this data. Also, the dataset shows a balanced representation of the different lung conditions including 5000 images in each class.


Images for lung_n category
Images for lung_scc category
Images for lung_aca category

Preprocessing Steps:

One of the important steps in this research is preprocessing, which is performed in different steps. After data extraction from the zip code file, we encoded each class into numerical forms. Label encoding is to convert categorical labels into numerical values, and we use label encoding because most machine learning algorithms require numerical inputs. Moreover, exploratory data analysis has been performed on data to find the distribution of images according to the three classes. The bar chart of the image's distribution is presented as below:

Class	*Number of Images*
lung_aca (Lung Adenocarcinomas	5000
lung_n(Normal Class)	5000
lung_scc(Lung Squamous Cell Carcinomas)	5000



Also, we have performed image resizing by changing the size of the images to a uniform dimension of 224x224 pixels. This has been done by using OpenCV library and Numpy library in python. This helps us to standardize the input dimension for the convolutional neural network (CNN) architecture. Next step is pixel values of the images normalization which has been performed by using a preprocess_input function provided by the TensorFlow framework.

In machine learning models, especially in classification tasks, we often have categorical labels like different classes of lung tissues - normal or cancerous. But Machine learning models work better with numeric data. So, in one-hot encoding, a binary variable with the values 0 or 1 is mapped to each category, and a new variable is created for each level of a categorical feature. [18] For example, if we have two classes, "Normal" and "Cancerous," one-hot encoding would represent "Normal" as [1, 0] and "Cancerous" as [0, 1]. Each class is assigned a unique binary representation. Furthermore, all the images are converted into the format we want.

The image size which is the constant that represents the desired size for our input images is 224 pixels. The split, which is the constant that determines the ratio in which our dataset will be split into training and validation sets is defined as 0.4. This split has been performed with a ratio of 60:40 to be sure that there is an adequate amount of data in both training and validation categories. So, we can say that the size of the training set is 9000 images, and the size of the validation set is 6000 images. The number of epochs which is defined as the many times the entire dataset will be passed forward and backward through the neural network during training is considered as 5 epochs. And the dataset is divided into batches during training, and each batch is processed by the neural network before updating the weights. This constant defines the number of samples in each batch. A smaller batch size often provides more frequent updates but might be computationally expensive. So, the batch size is 16 in this study. Then we use one hot encoding which transforms categorical labels into a binary matrix where each class is represented by a unique column, and each row corresponds to a sample. Then before splitting the data, we adjusted the normalization which helps us to be sure that both training and validation sets are processed consistently.

Model Development

Today, Convolutional Neural Networks (CNN) are very popular machine learning algorithms in performing image analysis. [7] One of the advantages of CNNs is that they maintain spatial relationships while filtering input images. [6] CNNs have been used in different areas including image classification, localization, detection, segmentation, and registration. We can use CNNs in processing both 2-dimensional images and 3-dimensional images with minor modifications.[6] CNNs and Recurrent Neural Networks (RNNs) are supervised machine learning algorithms, so we need to perform training data while using them. [6]

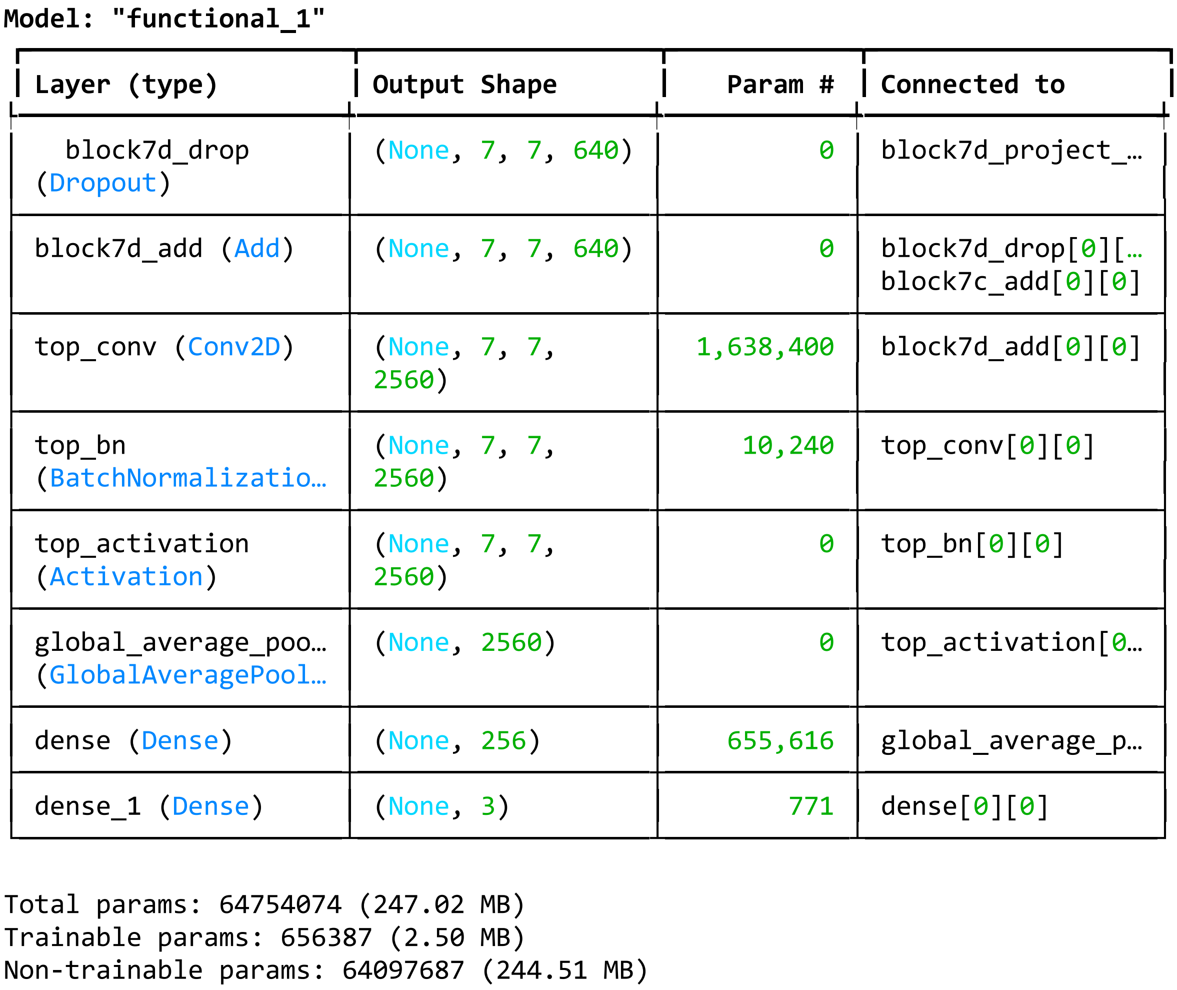
Convolution Layer

A convolution which is a fundamental operation in deep learning can be called an operation on two functions. One function represents the input values, such as pixel values, at a specific location in the image. The second function is a filter or kernel which is represented as an array of numbers. An output will be computed by using the dot product between the two functions. Then, the stride length determines the next location in the image where the filter is shifted. A feature map, also known as an activation map, is created by repeating the computation until the entire image is covered. The filter is activated on this map, and this is where a feature such as a straight line, a dot, or a curved edge can be seen. For example, when we put a face shot into a CNN, the filters first identify low-level features like edges and lines. Then, the feature maps become inputs for the next layer in the CNN architecture, and these build up to progressively higher features in successive layers, such a nose, eye, or ear. [6]

Model Architecture

In this study, the model architecture was implemented using the Functional API of Keras. We have implemented EfficientNetB7 architecture as the base model for feature extraction. EfficientNetB7 is a convolutional neural network (CNN) architecture which is very efficient in image classification. There are some key components of the model architecture that have been used in this model. A global average pooling layer is used to compute the average value of each feature map along its spatial dimensions (height and width). It replaces each 2D feature map with a single value, which is the average of all the values in that feature map. This operation results in a fixed-size output, where the length of the output vector is equal to the depth (number of channels) of the input feature maps, regardless of the input spatial dimensions. Also, fully connected layers (also known as dense layers) are added on top of the feature representation obtained from the pre-trained base model. Then, the final output layer of the model produces soft probabilities for the target classes using a SoftMax activation function. These probabilities represent the likelihood of each class, allowing for multi-class classification of lung cancer.

Model Evaluation



```
Model: "functional_1"
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┓
┃ Layer (type)        ┃ Output Shape      ┃    Param # ┃ Connected to      ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━┩
│   block7d_drop      │ (None, 7, 7, 640) │          0 │ block7d_project_… │
│ (Dropout)           │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ block7d_add (Add)   │ (None, 7, 7, 640) │          0 │ block7d_drop[0][… │
│                     │                   │            │ block7c_add[0][0] │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ top_conv (Conv2D)   │ (None, 7, 7,      │  1,638,400 │ block7d_add[0][0] │
│                     │ 2560)             │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ top_bn              │ (None, 7, 7,      │     10,240 │ top_conv[0][0]    │
│ (BatchNormalizatio… │ 2560)             │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ top_activation      │ (None, 7, 7,      │          0 │ top_bn[0][0]      │
│ (Activation)        │ 2560)             │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ global_average_poo… │ (None, 2560)      │          0 │ top_activation[0… │
│ (GlobalAveragePool… │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ dense (Dense)       │ (None, 256)       │    655,616 │ global_average_p… │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ dense_1 (Dense)     │ (None, 3)         │        771 │ dense[0][0]       │
└─────────────────────┴───────────────────┴────────────┴───────────────────┘

Total params: 64754074 (247.02 MB)
Trainable params: 656387 (2.50 MB)
Non-trainable params: 64097687 (244.51 MB)
```

In the process of training deep learning models, callbacks play a crucial role in monitoring and controlling the training process. A callback is a collection of operations that are used at specific points during the training process. Callbacks are a useful tool for viewing the internal states and statistics of the model while it is being trained [17]. We use callback to monitor the model improvement during each epoch. So, an early stopping is employed to prevent overfitting after optimizing the model using the Adam optimizer. We put callback as 90% which mean the algorthms stops when it reaches the best performance to prevent overfitting. And finally, the model performance is evaluated using evaluation metrics such as accuracy, precision, recall, and F1-score.

```
Epoch 1/5

563/563 [==============================] - ETA: 0s - loss: 0.1579 - accura
cy: 0.9362
Validation accuracy has reached 90%, so stopping further training.

563/563 [==============================] - 1864s 3s/step - loss: 0.1579 -
accuracy: 0.9362 - val_loss: 0.0794 - val_accuracy: 0.9673 - lr: 0.0010
```

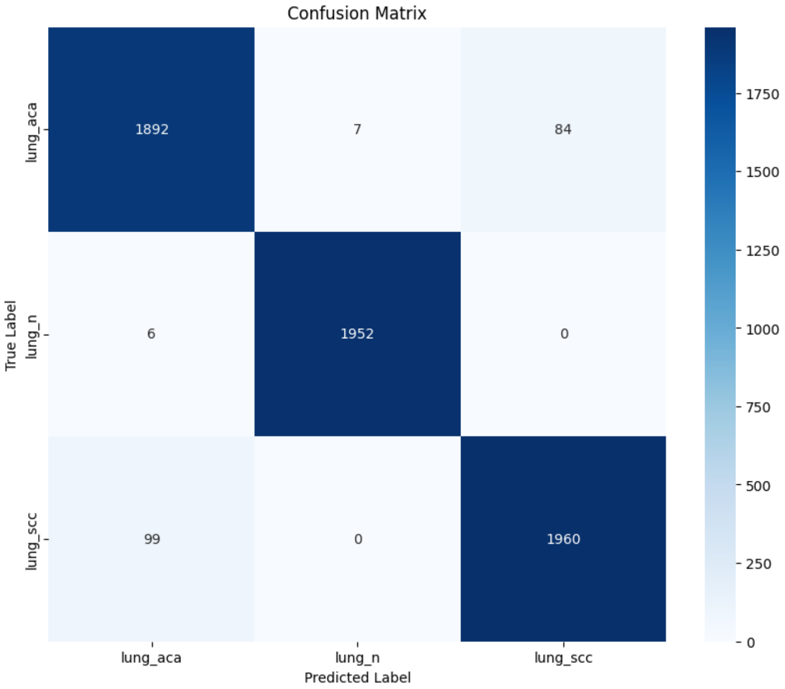
Experimental Results

Training and Validation Accuracy are two important metrics for evaluating the model’s performance. Training accuracy represents the percentage of instances in the training dataset that are classified correctly. When the model iterates through each epoch of training, the model learns from the data, but we should be careful because the model might overfit the data by using too many epochs. On the other hand, we have a separate validation dataset which is not used during training. So, we use validation accuracy to measure the model's performance on this part of the data. In this training process, the final training accuracy was approximately 93% and a training loss was 0.1579. Also, the validation accuracy reached a high value of 96% with a validation loss of 0.0794.

Our classification results, as shown in the classification report, underscore the performance of our model across all classes, including lung_scc, lung_aca, and lung_n. As can be seen from the results, the precision, recall, and F1-scores of our model demonstrate almost perfect performance. Achieving an overall accuracy of 97% on the validation dataset shows the robustness and efficacy of our proposed model. These results suggest that our model has a significant ability to diagnose and classify lung cancer which leads us to accurate patient care decisions.

```
                precision    recall   f1-score   support

    lung_aca        0.95       0.95       0.95      1983
      lung_n        1.00       1.00       1.00      1958
    lung_scc        0.96       0.95       0.96      2059
    accuracy                           0.97      6000
   macro avg       0.97      0.97      0.97      6000
weighted avg       0.97      0.97      0.97      6000
```



Discussion

The results of our study demonstrate the great performance of our deep learning model in predicting lung cancer. The classification report shows high precision, recall, and F1-score values across all classes, including lung_scc, lung_aca, and lung_n. These metrics indicate the model's ability to accurately classify instances of each category. The overall accuracy of 97% on the validation dataset emphasizes the robustness of our model in distinguishing between different lung cancer categories.

The success of our model can be because of several factors. Firstly, the utilization of transfer learning with the EfficientNetB7 architecture enables our model to leverage knowledge learned from a large-scale dataset (ImageNet) to effectively extract relevant features from medical images. Also, the global average pooling and dense layers are used to perform the transformation of extracted features into meaningful predictions, enhancing the model's performance. Furthermore, early stopping regularization helps prevent overfitting, ensuring that our model generalizes well to unseen data.

Conclusion

In conclusion, our study shows the good performance of deep learning for lung cancer prediction using medical imaging data. The high accuracy and performance of our model, as shown by the classification results, is a reliable tool for assisting healthcare professionals in diagnosing and classifying lung cancer cases. By providing accurate and timely predictions, our model has the potential to be used in helping the patients to have proper treatment strategies.

References

[1] Don S. Dizon, Arif H. Kamal, Cancer statistics 2024: All hands on deck, CA: A Cancer Journal for Clinicians, 10.3322/caac.21824, 74, 1, (8-9), (2024).

[2] 34 Islami F, Sauer AG, Miller KD, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable factors in the United States. CA Cancer J Clin. 2018; 68(1): 31-54. doi:10.3322/caac.21440

[3] 88 Kratzer TB, Bandi P, Freedman ND, et al. Lung cancer statistics, 2023. Cancer. 2023; [In press].

[4] Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol. 2020;1213:3-21. doi: 10.1007/978-3-030-33128-3_1. PMID: 32030660; PMCID: PMC7442218.

[5] 8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. [PubMed] [Google Scholar]

[6] J. Ker, L. Wang, J. Rao and T. Lim, "Deep Learning Applications in Medical Image Analysis," in IEEE Access, vol. 6, pp. 9375-9389, 2018, doi: 10.1109/ACCESS.2017.2788044.

[7] 4. G. Litjens et al., A survey on deep learning in medical image analysis, Jun. 2017, [online] Available: https://arxiv.org/abs/1702.05747.

[8] Zhang C., Sun X., Guo X., Zhang X., Yang X., Wu Y., Zhong W. Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network. Oncologist. 2019;24:1159–1165.

[9] Zhang C., Sun X., Guo X., Zhang X., Yang X., Wu Y., Zhong W. Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network. Oncologist. 2019;24:1159–1165.

[10] Nasser I.M., Naser A. Lung cancer detection using artificial neural network. Int. J. Eng. Inf. Syst. 2019;3:17–23.

[11] Cifci M.A. SegChaNet: A Novel Model for Lung Cancer Segmentation in CT Scans. Appl. Bionics Biomech. 2022;2022:1139587. doi: 10.1155/2022/1139587.

[12] Jakimovski G., Davcev D. Using Double Convolution Neural Network for Lung Cancer Stage Detection. Appl. Sci. 2019;9:427. doi: 10.3390/app9030427

[13] Wang S., Shi J., Ye Z., Dong D., Yu D., Zhou M., Liu Y., Gevaert O., Wang K., Zhu Y., et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur. Respir. J. 2019;53:1800986.

[14] Wang C., Shao J., Lv J., Cao Y., Zhu C., Li J., Shen W., Shi L., Liu D., Li W. Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography. Transl. Oncol. 2021;14:101141. doi: 10.1016/j.tranon.2021.101141.

[15] R. Kaur, R. Kumar and M. Gupta, "Review on Transfer Learning for Convolutional Neural Network," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 922-926, doi: 10.1109/ICAC3N53548.2021.9725474.

[16] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images", Int. J. Sci. Res. Publ., vol. 9, no. 10, pp. p9420, 2019.

[17] Chollet, F., et al. (2022). Keras Documentation. TensorFlow. Retrieved from https://keras.io/api/callbacks/

[18] B. Roy, Type of Categorical Encoding, 2020, [online] Available: https://www.analyticsvidhya.com/blog/2020/08/types-of-categoricaldata-encoding/.