Prediction of Lung Cancer using VGG19 Transfer Learning

Mrs K Rashmi[1], B Sai Shashank[2], B Pranaya[3], R Nikhitha Reddy[4], K Anvitha Rao[5]

[1]AssistantProfessor, Computer Science and Engineering, Anurag University
[2]Student, Computer Science and Engineering, Anurag University
[3]Student, Computer Science and Engineering, Anurag University
[4]Student, Computer Science and Engineering, Anurag University
[5]Student, Computer Science and Engineering, Anurag University

Abstract

Lung cancer is one of the deadly diseases whose prediction is required to reduce the death rate. So, Artificial intelligence is used on CT scan images are used for achieve better accuracy in an automated manner. Deep Learning is one of the emerging trends for predicting values.

Convolution Neural Networks is one of the deep learning algorithms implemented to sample produces better outcomes than other machine learning algorithms. In this paper, the dataset has been taken with 1000 images of chest scans for different types of lung cancers such as Adenocarcinoma, Benign, and Squamous Cell Carcinoma. Multiple machine learning algorithms have been compared and then it has been confirmed that CNN is one of the best among all to check the accuracy of the prediction.

The existing system includes VGG-16 but the model achieves only 77.62% which is not very effective, so, the proposed system implements the VGG-19 transfer learning model on datasets with different types of lung cancer, thus helping to check the severity and precautions for the same in a distinct manner.

1. Introduction

The abnormal growth of cells in the human Lung is called Lung Cancer. Lung cancer is one of the most serious diseases in the world today, and it has been the leading cause of mortality in the previous several decades. It also kills more people each year than breast, prostate, and colon cancer put together. The addiction to cigarettes is one of the leading causes of lung cancer. Furthermore, carcinogenic surroundings such as radioactive gas and air pollution contribute to the spread of this disease. In addition, genetic factors also have a major contribution to lung cancer. Uncontrolled magnification of tissue creates lung cancer. Primary originate from cells within secondary cancer begin in another part of the body and therefore spread to the lungs. In the human body, there are two types of cells. Normal cells are small and confined, whereas cancer-affected cells are rapidly forming and can be easily spotted. These cells appear to be aberrant and dissimilar to regular cells. This type of cell grows quickly and is more prone to spread.

Of all the diseases that have existed in mankind lung cancer has emerged as one of the most fatal one. Also, it is one of the most common and contributing to deaths among all cancers. Cases of lung cancer are increasing rapidly. There are about 70,000 cases per year in India. The disease tends to be asymptomatic mostly in its earlier stages thus making it nearly impossible to detect. That's why early cancer detection plays an important part in saving lives. Early detection can give a patient a better chance to be cured and recover. Technology plays a major role in detecting cancer efficiently. Many researchers have proposed different methods based on their studies. In recent times, to use computer technology to solve this problem, several computer-aided diagnoses (CAD) techniques as well as systems have been proposed, developed as well as emerged. Those systems use various Machine learning techniques as well as deep learning techniques, there also have been several methods based on image processing-based techniques to predict the malignancy level of cancer. Here, in this system, the aim will be to analyze image datasets with transfer learning techniques to classify and detect lung cancer in its early stages.

2. Literature Survey

Multiple researchers surveyed and resulted to some outcomes such as C. Yao et al. [11] designed a CNN model taking self-made data set achieved 90% accuracy but the validity of data and accuracy on real time data sets remained doubtful. So, the method needs to be tested at several other data sets to check its reliability. In the same manner M. Norouzi [1] claimed that nanotechnology is one of the efficient methods to be applied for therapies for lung cancer rather than complicated chemotherapies and it can also reduce the amount of toxicity. The survey has proved nanotechnology a great method to be considered with conditions applied because patient selection and combinations of multiple treatment provides an advanced enhancement.

Multiple papers have been surveyed in J. Wang et al. [9]and verified that false positive rates of those are quite high which decrease the accuracy, so to overcome the problem, 3 dimensional - convolutional neural network, VGG 16, Alex Net and Multi Crop Net L.ye et al. [12] advanced from 2-dimensional architecture which extracted 8.28% which is relatively low when compared with other algorithms.

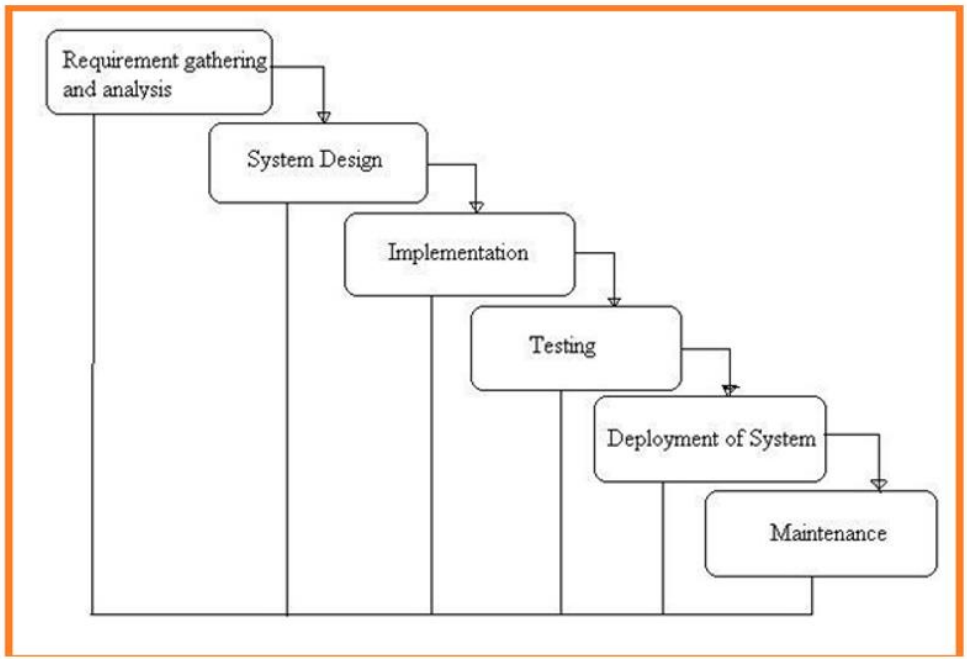
A data set of 1000+ images has been taken in J. Wang et al. [13] with stage T1a-3N0M0, NSLC where it has been claimed that non-small lung cancer is more dangerous than small lung cancer, the death rate compared, the former is considered more harmful.

Survival rate has been checked by A. Agaimy et al. [6] after the first diagnosis of non-small lung cancer to check the criticality of it. The data contains both the genders which includes 8 males and 6 females under the age group of 52to 85 years (median 60). The maximum survival rate after the diagnosis was of one patient who was having cerebral metastasis, alive even after 45 months.

VGG 16 implementation is also done in paper S. T. M. Sheriff et al. [14] where the result only reveals the presence of lung cancer or not with not so good accuracy, so the inspiration has been taken from the paper to implement not just on cancerous images but also the types of lung cancer so as to check that how critical the consequences would be and what precautions need to take care.

Multiple CNN models has been checked in M. Phankokkruad et al. [15] such as VGG16, ResNet50V2, and DenseNet201 which are based on transfer learning. Every model predicted its accuracy provided as 62%, 90%, and 89%, respectively.

System Analysis



What is Waterfall Model?
Waterfall Model is a sequential model that divides software development into different phases. Each phase is designed for performing specific activity during SDLC phase. It was introduced in 1970 by Winston Royce.

Requirements:
The first phase involves understanding what needs to design and what is its function, purpose, etc. Here, the specifications of the input and output or the final product are studied and marked.

System Design:
The requirement specifications from the first phase are studied in this phase and system design is prepared. System Design helps in specifying hardware and system requirements and also helps in defining overall system architecture. The software code to be written in the next stage is created now.

Implementation:
With inputs from system design, the system is first developed in small programs called units, which are integrated into the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.

Integration and Testing:
All the units developed in the implementation phase are integrated into a system after testing of each unit. The software designed, needs to go through constant software testing to find out if there are any flaws or errors. Testing is done so that the client does not face any problem during the installation of the software.

Deployment of System:
Once the functional and non-functional testing is done, the product is deployed in the customer environment or released into the market.

Maintenance:
This step occurs after installation, and involves making modifications to the system or an individual component to alter attributes or improve performance. These modifications arise either due to change requests initiated by the customer, or defects uncovered during live use of the system. The client is provided with regular maintenance and support for the developed software.

3. Methodology

VGG-19:

VGG-19 is a 19-layer deep convolutional neural network (CNN). It is a popular method for image classification because it uses multiple 3 × 3 filters in each convolutional layer. VGG-19 is trained on the ImageNet database, which contains a million images of 1000 categories. A pre-trained version of the network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. Here VGG-19 can achieve an accuracy of 95% with a loss of 17%. Compared with existing methods, VGG-19 has a faster training speed, fewer training samples per time, and higher accuracy.

The VGG network is constructed with very small convolutional filters. The VGG-19 consists of 16 convolutional layers and three fully connected layers.

Let’s take a brief look at the architecture of VGG:

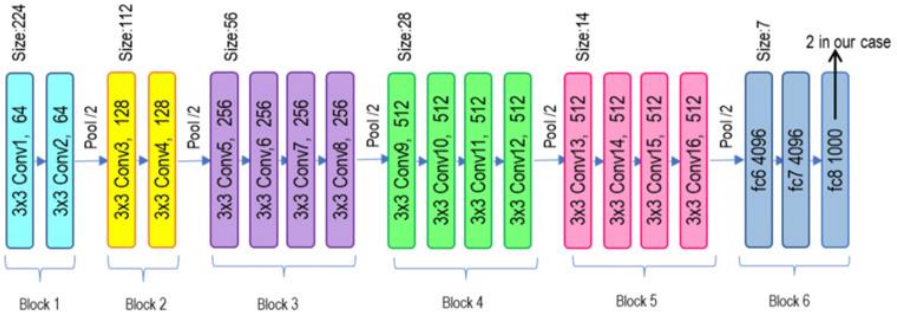
Input: The VGG Net takes in an image input size of 224×224. For the ImageNet competition, the creators of the model cropped out the center 224×224 patch in each image to keep the input size of the image consistent.

Convolutional Layers: VGG's convolutional layers leverage a minimal receptive field, i.e., 3×3, the smallest possible size that still captures up/down and left/right. Moreover, there are also 1×1 convolution filters acting as a linear transformation of the input.

This is followed by a Re LU unit, which is a huge innovation from Alex Net that reduces training time. Re LU stands for rectified linear unit activation function; it is a piecewise linear function that will output the input if positive; otherwise, the output is zero. The convolution stride is fixed at 1 pixel to keep the spatial resolution preserved after convolution (stride is the number of pixels shifts over the input matrix).

Hidden Layers: All the hidden layers in the VGG network use Re LU. VGG does not usually leverage Local Response Normalization (LRN) as it increases memory consumption and training time. Moreover, it makes no improvements to overall accuracy.

Fully-Connected Layers: The VGG Net has three fully connected layers. Out of the three layers, the first two have 4096 channels each, and the third has 1000 channels, 1 for each class.



VGG-19 has 16 convolution layers grouped into 5 blocks. After every block, there is a Max pool layer that decreases the size of the input image by 2 and increases the number of filters of the convolution layer also by 2. The dimensions of the last three dense layers in block 6 are 4096, 4096, and 1000 respectively. VGG classifies the input images into 1000 different categories. As there are two output classes in this study the dimension of fc8 is set to two.

3.1. Dataset Collection:

In this proposed system, we are acquiring the Lung Cancer Histopathological Images dataset from the Kaggle web repository. This dataset contains 3000 histopathological images with 3 classes like Adenocarcinoma, Benign, and Squamous-Carcinoma. Each class has 1000 images and all images are 768 x 768 pixels in size and are in jpeg file format.

3.2. Image Pre-processing:

The dataset contains folder structures with images, so it needs to read the images from that directory with Python libraries *OS* and *path*. The machines cannot understand images directly; therefore, it is mandatory to convert the images into pixel format with the Python library Numpy to get features from images. Later the image dataset will be separated into independent and dependent features. Here independent features are like image pixels which are stored in a list and disease names or classed or target values are treated as dependent values which they can also store in a separate list.

3.3. Split Dataset:

Based on values of independent features and target features, the dataset will be split with 70 to 30 ratios. Here 70 percent indicates the training set and 30 percent indicates the testing set.

3.4. Training the Model:

Here the CNN architecture will be created based on the VGG-19 model which has 16 input layers and 3 output layers. Thereafter, the VGG-19 model will be trained with a training set and generated the training model which will be used for the prediction of lung cancer further.

3.5. Performance Evaluations

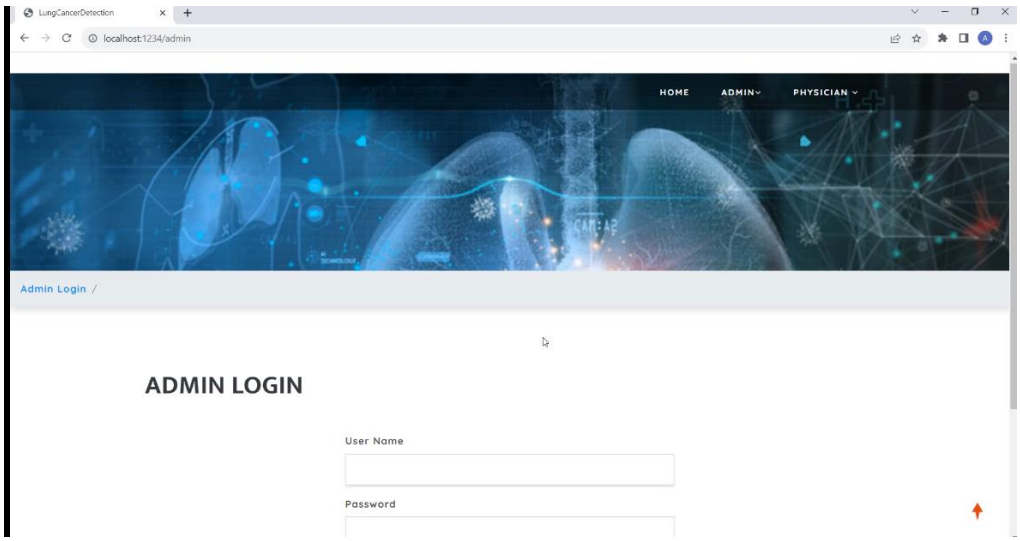
The performance evaluations will be generated with the help of testing the image dataset. Here the trained CNN model will used to calculate performance metrics such as accuracy, loss, precision, and recall by taking the input of the testing dataset. For visualization, the line chart graph will be plotted with the performance metrics of the deep learning model.

3.6. Lung Cancer Detection

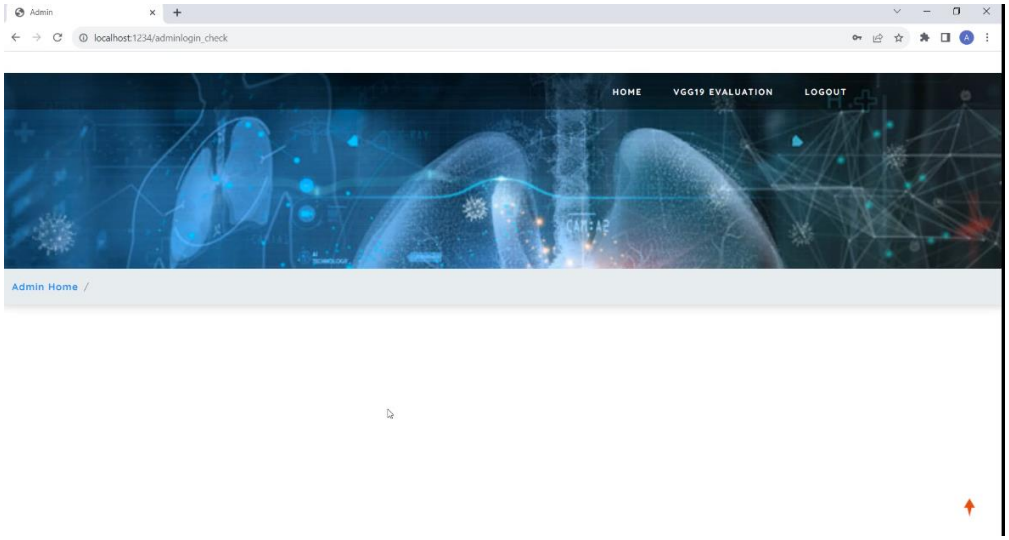
In this stage, it has to browse the Pathology image as an input image to detect lung cancer. Here from this testing input image, this system extracts the features of the image and feeds these features to the deep learning model CNN then this predictor model can classify the lung cancer disease.

4. Results

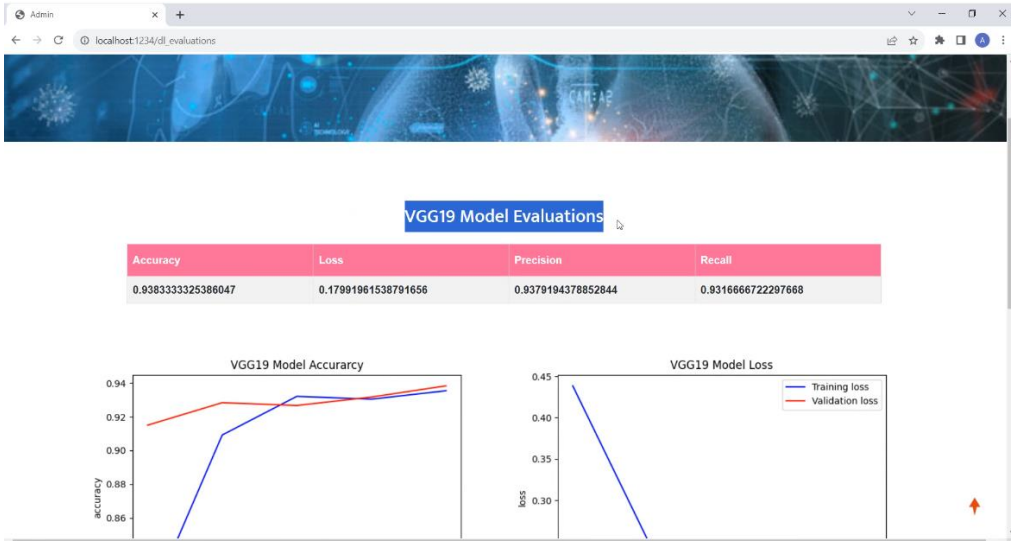
Home Screen



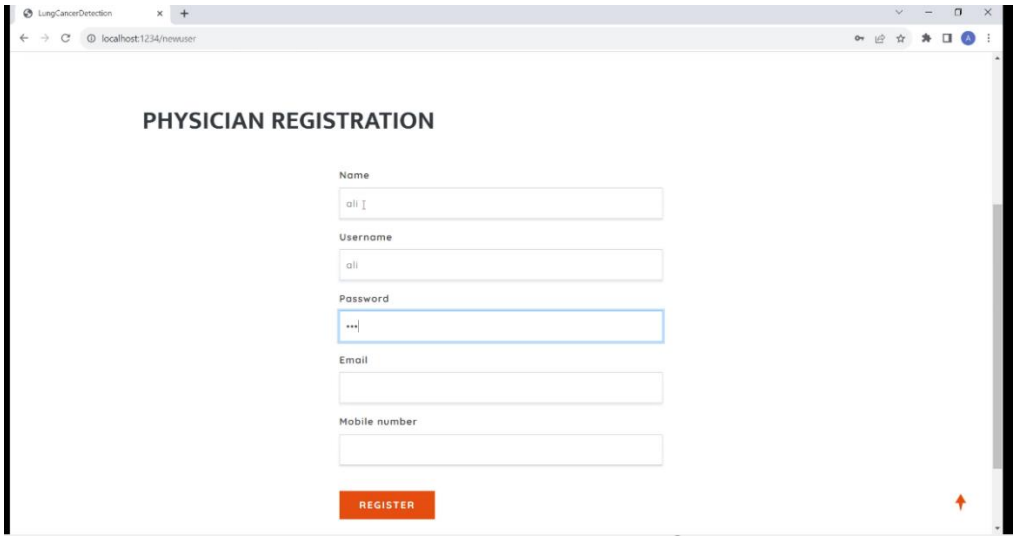
Admin Login



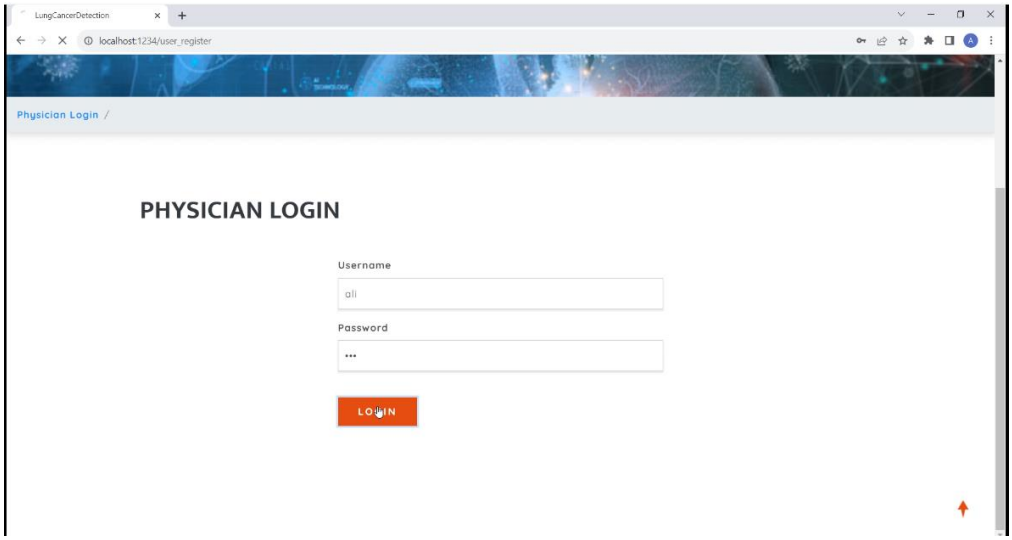
Admin Home



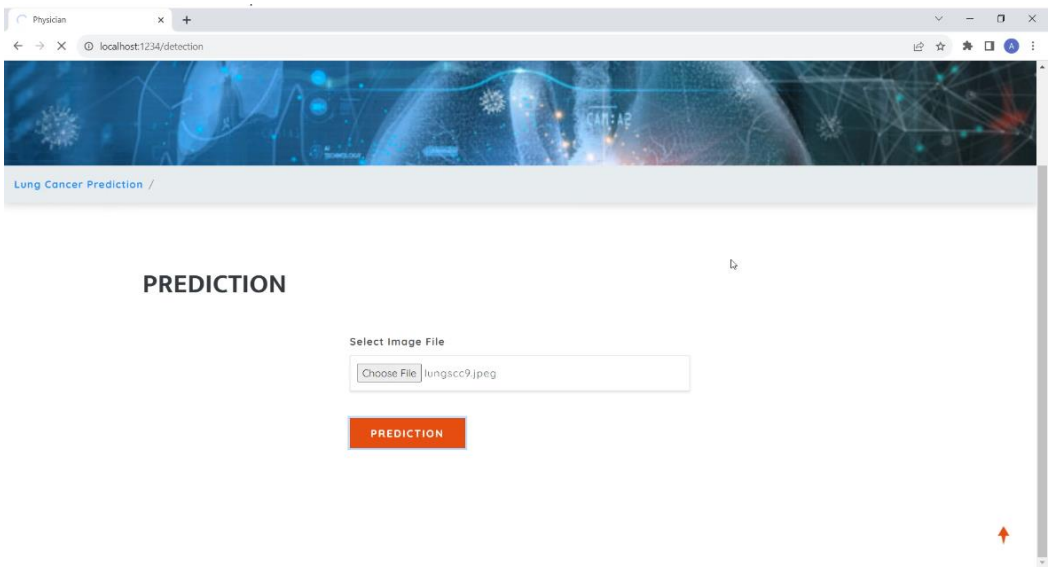
VGG-19 Model Evaluations



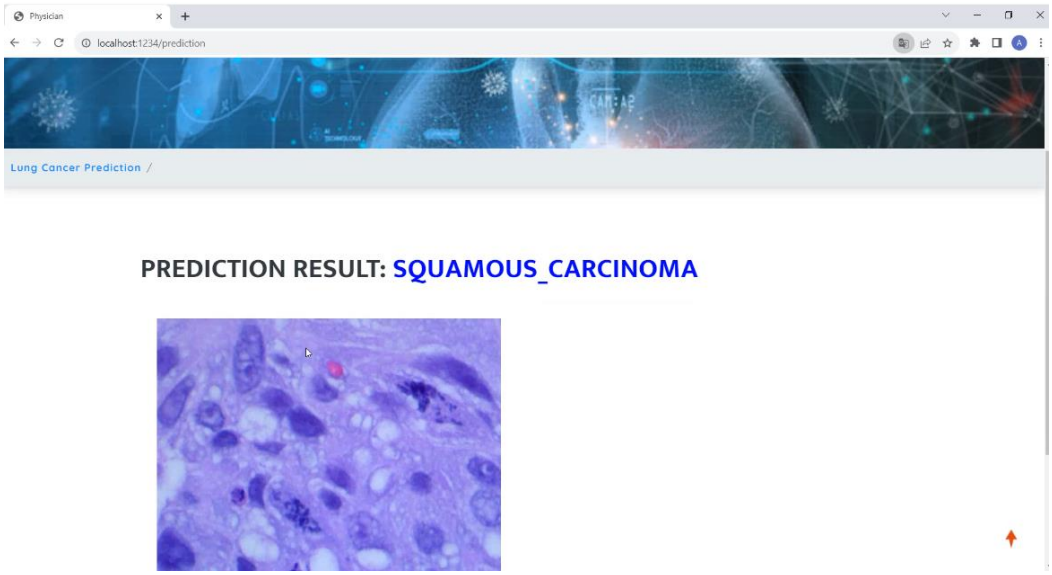
User registration



User Login



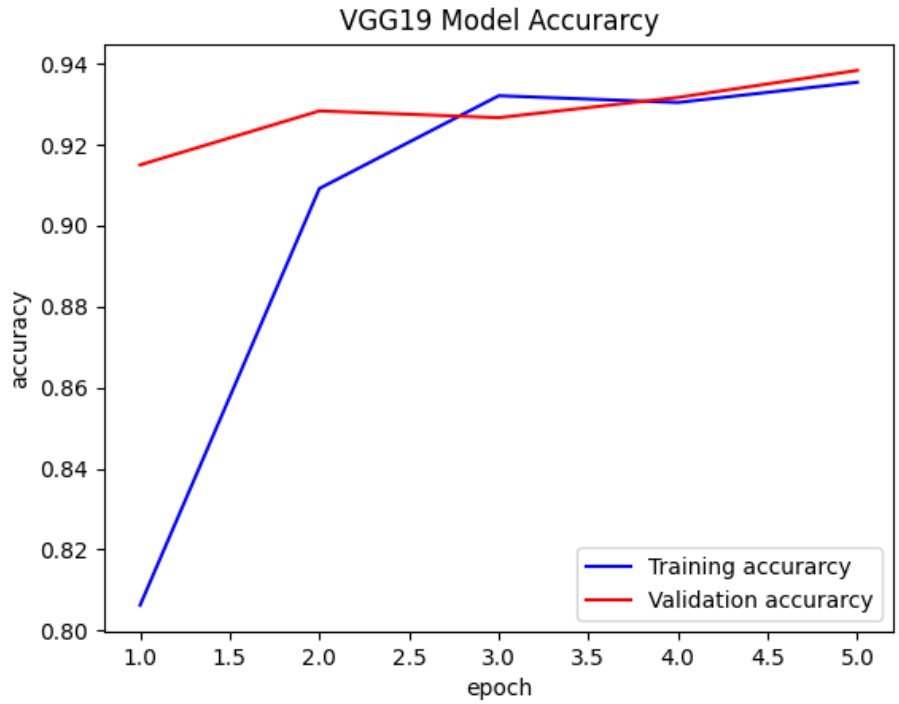
Prediction page



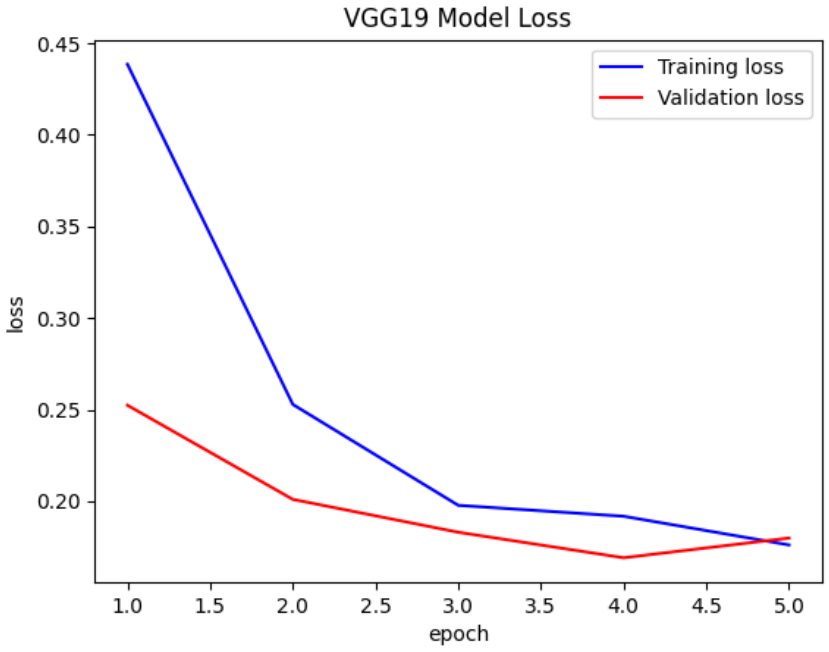
Test Results

Graphs in Application

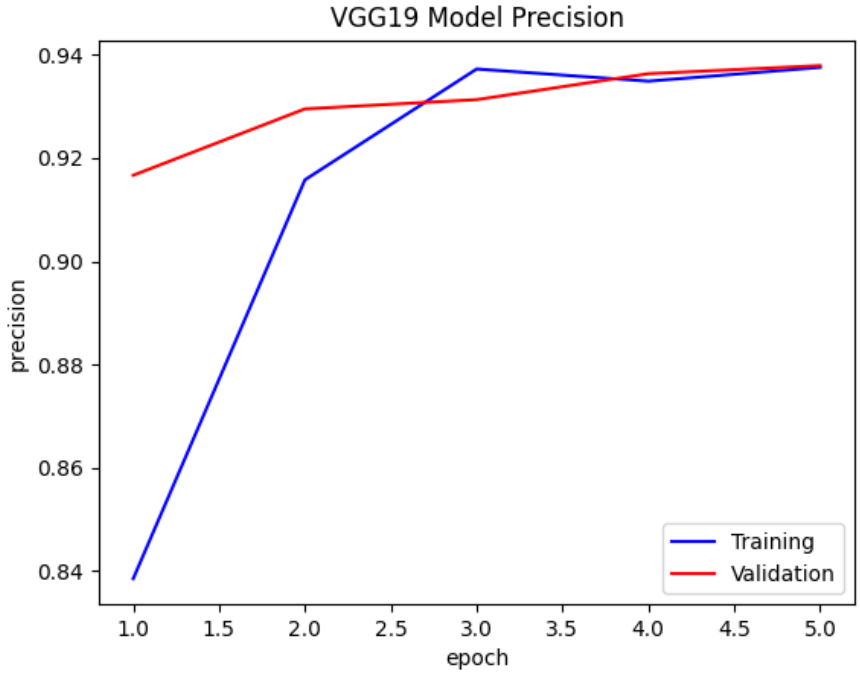
Accuracy



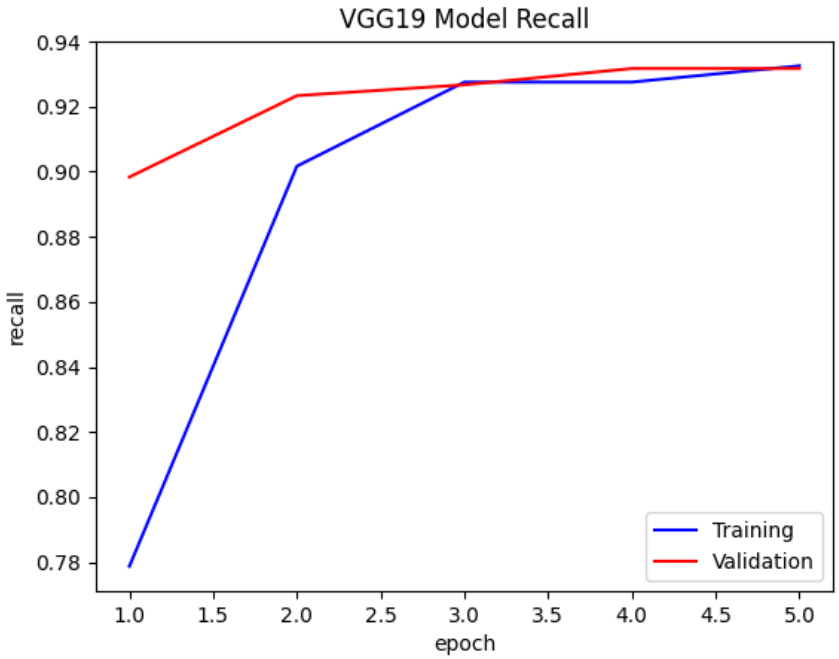
Loss



Precision


VGG19 Model Precision
precision
epoch
Training
Validation
0.94
0.92
0.90
0.88
0.86
0.84
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Recall


VGG19 Model Recall
recall
epoch
Training
Validation
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

5. Conclusion

We used a pre-trained system (VGG19) to analyse lung scans and predict cancer. Our system achieved an accuracy of (91%) in identifying lung cancer.

This method is helpful because it saves time and works well. There's room for improvement, like fine-tuning the system and using more data. Overall, this approach shows promise for helping doctors find lung cancer.

The data-set when applied VGG-16 algorithm to check the accuracy and other metrics, resulted in the provided outcomes: The best result predicted with the model is provided. The figure 5 predicts accuracy and loss, figure 6 predicts AUC and precision and figure 7 predicts precision and F1 score. The uphill and downfall has been shown in the figures provided.

6. 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 in Oncology, 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, et al., “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, “Computational detection of a genome instability-derived lncrna signaturefor 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 lung adenocarcinoma,” 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 m olecular analysis of 14 cases indicates an unusual pattern of dedifferentiated metastases,” Virchows Archiv, pp. 1–11, 2021.

[7] K. I. Tosios, V. Papanikolaou, D. Vlachodimitropoulos, 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,” Physics inMedicine & Biology, vol. 64, no. 17, p. 175012, 2019.

[10] images from biorender, “www.biorender.com,” 2022.