ZKG INTERNATIONAL ISSN: 2366-1313

LUNG CANCER PREDICTION USING BIGDATA

[1]DR M.VINAYA BABU, [2]ANIREDDY SOWMYA, [3]ABHISHEK KUMAR THAKUR, [4]ATKAPURAM SAITEJA

[1](Assistant Professor) ,**CSE.** Teegala Krishna Reddy Engineering College Hyderabad

[234]B,tech scholar ,**CSE.** Teegala Krishna Reddy Engineering College Hyderabad

ABSTRACT

Lung Cancer is one of the leading lives taking cancer worldwide. Early detection and treatment are crucial for patient recovery. Medical professionals use histopathological images of biopsied tissue from potentially infected areas of lungs for diagnosis. Most of the time, the diagnosis regarding the types of lung cancer is error-prone and time-consuming. Convolutional Neural networks can identify and classify lung cancer types with greater accuracy in a shorter period, which is crucial for determining patients' right treatment procedure and their survival rate. Benign tissue, Adenocarcinoma, and squamous cell carcinoma are considered in this research work.

1. INTRODUCTION

1.1 INTRODUCTION

Lung cancer is prominent cancer among both men and women, making up almost 25% of all cancer deaths [1]. The primary cause of death from lung cancer, about 80% is from smoking. Lung cancer in non-smokers can be caused by exposure to radon, second-hand smoke, air pollution, or other factors like workplace exposures to asbestos, diesel exhaust, or certain other chemicals lung cancers some people who do not smoke [2]. Various tests like imaging sets (x-ray, CT scan), Sputum cytology, and tissue sampling (biopsy) are carried out to look for cancerous cells and rule out other possible conditions. While performing the biopsy, evaluation of the microscopic histopathology slides by experienced pathologists is indispensable to establishing the diagnosis [3], [4], [5], and defines the types and subtypes of lung cancers [6]. For pathologists and other medical professionals diagnosing lung cancer and the types is a time-consuming process. There is a significant change the cancer types are misdiagnosed, which directs to incorrect treatment and may cost patients' lives. Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience [7][8]. In previous research papers, most of the authors considered using x-rays, CT scans images with machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN), and Convolutional Neural Network (CNN) for lung cancer detection and recognition purpose. Some papers also considered using histopathological images, but they distinguish between carcinomas and non-carcinomas images and with lower accuracy. This research paper has considered using Convolutional Neural Network (CNN) architecture to classify the benign, Adenocarcinoma, and squamous cell carcinomas. We have not found other papers using the CNN model to classify only the given three different histopathological images and the given model's accuracy.

1.1 Problem statement

Computer-aided diagnosis (CAD) is cuttingedge technology in the field of medicine that interfaces computer science and medicine. CAD systems imitate the skilled human expert to make diagnostic decisions with the help of diagnostic rules. The performance of CAD systems can improve over time and advanced CAD can infer new knowledge by analysing the clinical data. To learn such capability the system must have a feedback mechanism where the learning happens by successes and failures. During the last century, there is a dramatic improvement in human expertise and examination tools such as X-ray, MRI, CT, and ultrasound. Different image processing methods have been innovated for detecting cancer and Implemented as a median-wiener filter in the preprocessing step. To detect whether the nodule is cancerous or not, the classification network has been used, such as Support Vector Machines (SVM).

DISADVANTAGES OF EXISTING SYSTEM:

Most of the authors considered using x-rays, CT scans images with machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN), for lung cancer detection and recognition purpose. Some papers also considered using histopathological images, but they distinguish between carcinomas and non- carcinomas images and with lower accuracy.

1.2 Objective

The objective of lung cancer prediction using big data is to develop and implement advanced computational models and data analytics techniques to accurately identify individuals at high risk of developing lung cancer. By analyzing large-scale datasets encompassing diverse factors such as genetic predisposition, environmental exposures, lifestyle habits, and medical history, the aim is to create predictive models that can assist in early detection, personalized treatment planning, and ultimately, improve patient outcomes and survival rates.

- This project has considered using Convolutional Neural Network (CNN) architecture to classify the benign, Adenocarcinoma, and squamous cell carcinomas. We have implemented this project using the CNN VGG 19 model to classify only the given three different histopathological images and the given model's accuracy.

- A convolutional neural network (CNN) was implemented to classify an image of three different categories benign, Adenocarcinoma, and squamous cell carcinoma.

- The model was able to achieve higher accuracy.

2. LITERATURE SURVEY

2.1 LITERATURE REVIEW

he authors W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn [9] used deep learning with a transfer learning approach to predict lung cancer from the chest X-ray images obtained from different data sources. Image size of 224X224 with 121-layer Densely Connected Convolutional Network (DenseNet-121) and a single sigmoid node was applied in a fully connected layer. The proposed model achieved 74.43±6.01% mean accuracy, 74.96±9.85% of mean specificity, and 74.68±15.33% mean sensitivity for different image source dataset. T. Atsushi, T. Tetsuya, K. Yuka, and F. Hiroshi [10] applied Deep Convolutional Neural Network (DCNN) on cytological images to automate lung cancer type classification. They considered Small cell carcinoma, Squamous cell carcinoma, Adenocarcinoma images in their dataset.

The DCNN architecture of 3 convolution and pooling layers and 2 fully connected layers with dropout 0f 0.5 were used. The model developed was able to achieve the overall accuracy of 71.1%, which is quite low. W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale [11] proposed using image processing and machine learning (Support Vector Machine) for lung cancer detection on computed tomography (CT) images. Image processing like grayscale conversion, noise reduction, and binarization was carried out. Features like area, perimeter, and eccentricity from the segmented image region of interest were fed to the support vector machine (SVM) model. M. Šarić, M. Russo, M. Stella, and M. Sikora [12] proposed CNN architectures implementing VGG and ResNet for lung cancer detection using whole side histopathology images, and the output was compared using the receiver operating characteristic (ROC) plot. Patch level accuracy of 0.7541 and 0.7205 was obtained for VGG16 and ResNet50 respectively which is quite low. The authors explained that the given models' low accuracy was due to large pattern diversity through different slides.

The authors S. Sasikala, M. Bharathi, B. R. Sowmiya [13], proposed using CNN on CT scan images to detect and classify lung cancer. They used MATLAB for their work and has two phases in training to extract valuable volumetric features from input data as the first phase and classification as the second phase. Their proposed system could classify the cancerous and non-cancerous cells with 96% accuracy. SRS Chakravarthy, R. Harikumar [14], used Co-Occurrence Matrix (GLCM) and chaotic crow search algorithm (CCSA) for feature selection on computed tomography (CT) and applied probabilistic neural network (PNN) of the classification task. They found that the PNN model build on CCSA features performed better with 90% accuracy.

3. SYSTEM DESIGN

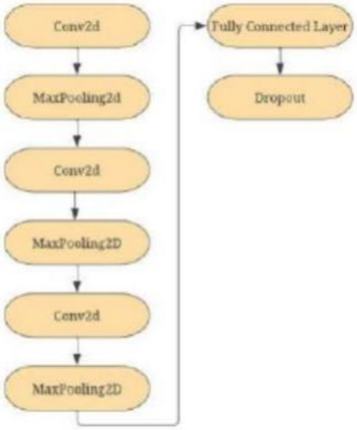
System Design Introduction:

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces. MODULES DATASET: The histopathology images are obtained from LC25000 Lung and colon histopathological image dataset [15]. Three classes of benign tissue, Adenocarcinoma, and squamous carcinoma cells of lungs with 5000 histopathology images in each category, are considered for our work.

PRE-PROCESSING: • Pre-processing is a procedure adopted to enhance the quality of images and increase visualization. In medical imaging, image processing is a crucial phase that helps to improve the images quality. This can be one of the most critical factors in achieving good results and accuracy in next phases of proposed methodology. Lung cancer images may contain a different issue that may lead to poor and low visualization of the image. If the images are poor or of low quality, it may lead to unsatisfactory results. During preprocessing phase.

DEEP LEANRING ALGORIHTM:

This study proposes three models, which are then used as comparisons. The first architecture used is the CNN model, as shown in Figure 3, using an input layer of 224x224 pixels. It can assist in image recognition and speed up the computation to distinguish four image classes consisting of Mild demented, Moderate demented, Non-demented, and Very mild demented. As shown in Figure 3, the first model uses three convolutional layers and three pooling layers by implementing max pooling with a 2 x 2 filter. This study also uses three convolutional layers with filters 64, 32, and 16, which use a 3 x 3 kernel and use relu activation. Furthermore, the fully connected layer is used, which has a flatten layer and a dense layer followed by a dropout layer 0.2 which uses relu activation. In the last process, a dense layer is used with softmax activation.



As shown in Figures 5, the second and third model architectures use Visual Geometry Group 16 (VGG16) and Visual Geometry Group 19 (VGG19) transfer learning.

3.1 SYSTEM ARCHITECTURE

A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system. Organized in a way that supports reasoning about the structures and behaviors of the system.

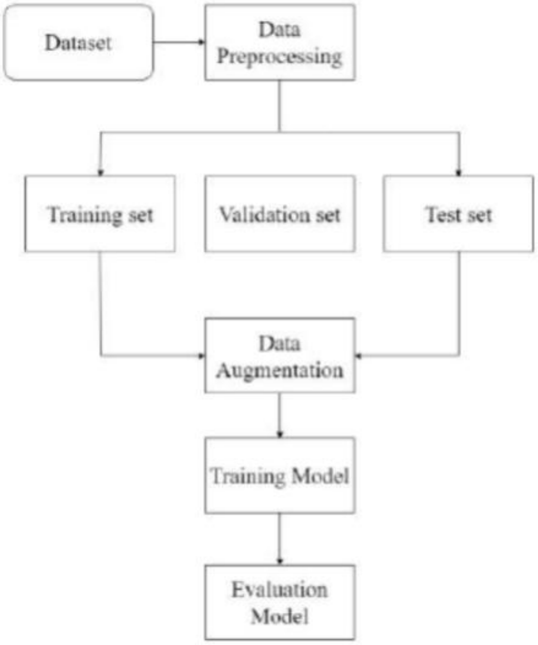


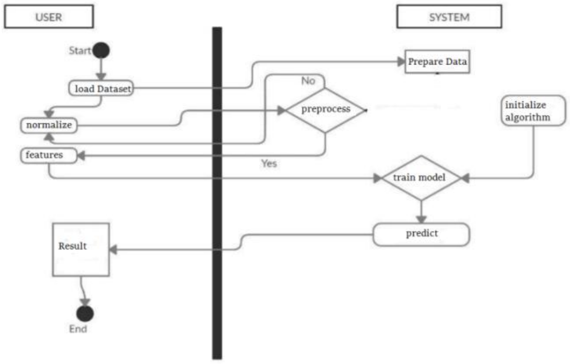
Figure 3. 1 System Architecture

Figure illustrates the research flow from start to finish. The research began by collecting datasets from the Kaggle website. The next stage is data preprocessing, such as rescaled image data and dividing the Lung cancer MRI dataset into three parts: training, test, and validation data. Data dividing aims to train the dataset so that the built model can learn according to the parameters built in the model.

Then the data is augmented to provide modifications to the lung cancer images. Furthermore, the model is trained using modified data so that the result of the model that has been built is the result of an evaluation of the model in the form of the value of each metric evaluated using model evaluate. Model evaluate is a function in the Keras library in the Python programming language that is useful for evaluating models trained using data validation or test data and appropriate labels.

3.2 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



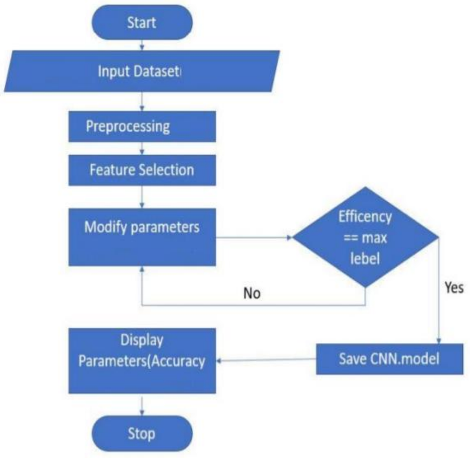
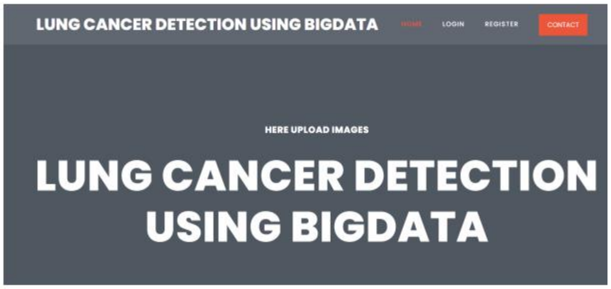
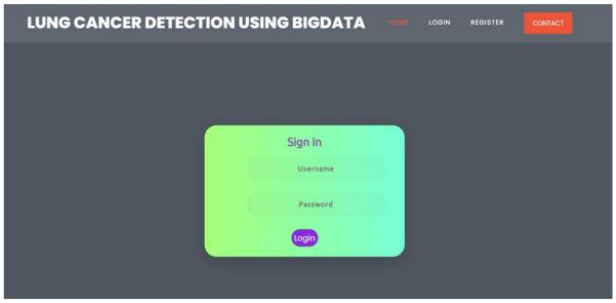


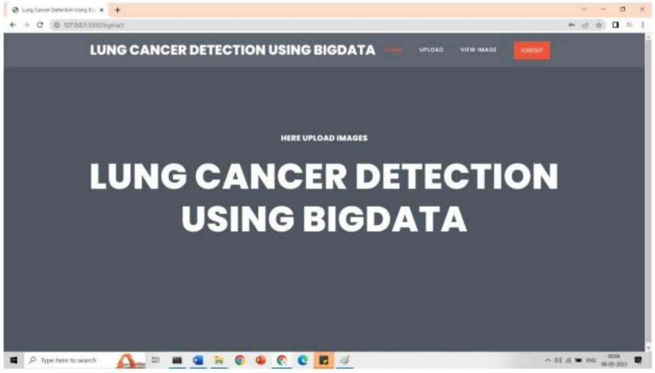
Figure 3.2 Activity Diagram

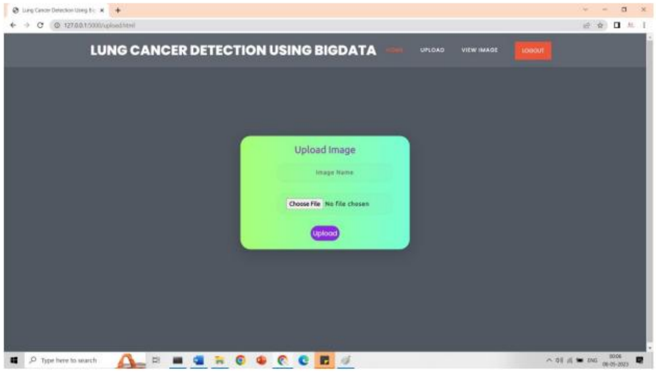
4. OUTPUT SCREENS

4.1 DATASET

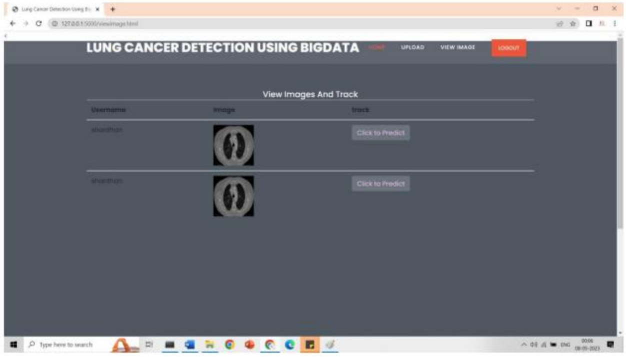

LUNG CANCER DETECTION USING BIGDATA
HOME
LOGIN
REGISTER
CONTACT
HERE UPLOAD IMAGES
LUNG CANCER DETECTION USING BIGDATA



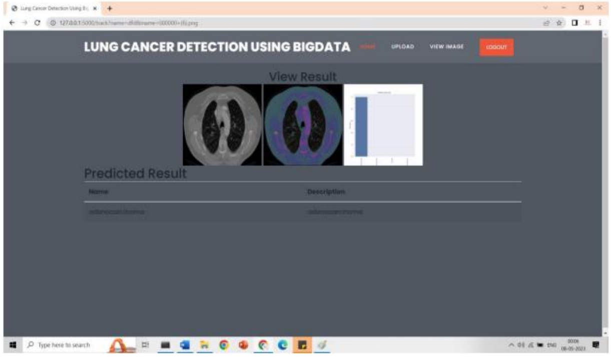

LUNG CANCER DETECTION USING BIGDATA
HOME
UPLOAD
VIEW IMAGE
LOGOUT
HERE UPLOAD IMAGES
LUNG CANCER DETECTION USING BIGDATA
Type here to search



4.2 UPLOAD TEST INFORMATION



ISSN: 2366-1313


Lung Cancer Detection Using B
LUNG CANCER DETECTION USING BIGDATA
HOME
UPLOAD
VIEW IMAGE
LOGOUT
Type here to search

View Result

Predicted Result

Name	Description
adenocarcinoma	adenocarcinoma

4.2 Prediction Of Cancer

5. CONCLUSION

Thisstudy proposes a transfer learning method to detect Lung cancer from structured image data. We performed several Lung cancer classifications using CNN, and VGG19 and proved that the method allows for multiple medical image classifications that can be applied to similar fields. This study applies several algorithms (CNN, VGG16, VGG19) for the multi-classification of Lung cancer datasets. The results of several medical image classifications are quite good, but there is still room for improvement. VGG19 gets the best performance with a value of 80% for accuracy, 60% for precision, 60%. On the other hand, got better performance results than the CNN handcraft model on all the performance metrics used. This study also has good medical image processing by using several evaluation metrics that are relevant to reveal the limited capacity of the model.

6. FUTURE ENHANCEMENT

The future scope of lung cancer prediction using Big Data models holds significant potential for advancements in several areas. In summary, employing CNN and VG199 algorithms to predict lung cancer has a bright future. Improved precision, personalized risk assessment, early identification, prognosis prediction, and real-time monitoring of lung cancer can all be attributed to ongoing research and development in these fields. The development of these models and their application to clinical practice, which will ultimately result in improved patient outcomes and efficient management of lung cancer, depend 27 heavily on collaborations between medical practitioners, data scientists, and researchers.

7. REFERENCES

- (2020) “American Cancer Society, Lung Cancer Statistics. [Online]”. Available: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

- [2] (2019) “American Cancer Society, Lung Cancer Causes. [Online].” Available: https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/what-causes.html

- [3] G. A. Silvestri, et al. "Noninvasive staging of non-small cell lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition)." Chest vol. 132, 3 Suppl (2007): 178S-201S. doi:10.1378/chest.07-1360.

- [4] W. D. Travis, et al. "International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma." Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer vol. 6, 2 (2011): 244-85. doi:10.1097/JTO.0b013e318206a221

- [5] L. G. Collins., C. Haines, R. Perkel & R. E. Enck. "Lung cancer: diagnosis and management." American family physician vol. 75, 1 (2007): 56-63.

- [6] K. Yu, C. Zhang, G. Berry, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nat Commun 7, 12474 (2016), doi: 10.1038/ncomms12474

- [7] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.