RESEARCH ARTICLE OPEN ACCESS

Lung Cancer Classification Using Deep Learning Coupled with Medical Imaging Techniques to Assist Early Diagnostics

Amaan Naseh[1], Md. Yusuf Azam[2], Zian Malik[3], Mohammad Rashid[4], Khyati Chopra[5]

[1](Automation and Robotics, Guru Gobind Singh Indraprastha University, Delhi, India, Email: naseh.amaan@gmail.com)
[2](Artificial Intelligence and Data Science, Guru Gobind Singh Indraprastha University, Delhi, India, Email: mdyusuf2521@gmail.com)
[3](Artificial Intelligence and Data Science, Guru Gobind Singh Indraprastha University, Delhi, India, Email: ziyanmalik321@gmail.com)
[4](Artificial Intelligence and Data Science, Guru Gobind Singh Indraprastha University, Delhi, India, Email: mohamma0rashid@gmail.com)
[5](Industrial Internet of Things, Guru Gobind Singh Indraprastha University, Delhi, India, Email: khyati.usar@ipu.ac.in)

Abstract

Cancer is oncogenic transformation of cells stimulated by carcinogens which stops contact inhibition that leads to tumors (benign or malign). As per World Health Organization (WHO), 10 million people died in 2020 due to cancer, most prominently by lung cancer, colon cancer, liver cancer, stomach cancer and breast cancer. Nearly, half of the cancer patients die due to late diagnosis. Current diagnostic techniques include biopsy and histopathology of tissues, radiography, computed tomography (CT), and magnetic resonance imaging (MRI). Some treatment techniques include chemotherapy, hormone therapy, and surgery. Although medical sector have these techniques, cancer at higher stages is still incurable, therefore late detection of cancer is fatal. With the onset of Industry 4.0, the era of Artificial Intelligence (AI) has been established. AI can be used to speed up the process of medical diagnosis, for example, convolutional neural network (CNN) to detect cancer based on medical report of the patient (X-ray, CT scan or MRI scan). In this research, we have developed two CNN models on MRI and histopathological images for lungs cancer diagnostics, with achieving validation accuracies as 96.45% and 99.57%, and validation losses as 0.12 and 0.01, respectively. A full stack website was developed by using flask as backend and React.js as frontend where CNN models were hosted on as backend API to serve image classification requests from user. The website was deployed using open-source deployment platforms.

Keywords *– Artificial Intelligence, Machine Learning, Neural Networks, Cancer Diagnostics, Medical Imaging Technique*

I. INTRODUCTION

Although the advancement in technology led to the enhancement of treatment of a variety of diseases including cancer, still cancer is fatal & incurable at higher stages. In most of the cases, late diagnosis of cancer increases the risk of survival of patient. Therefore, cancer diagnostic technique plays a crucial role in treatment & survival of patients. Conventional diagnostic techniques such as Medical Imaging including Radiography, Mammography, Computer-Aided Detection (CAD), etc. are time consuming as well as less accurate, due to lot of dependences & constraints. Traditional radiography techniques depends majorly on 'semantic features' (qualitative & quantitative analysis of phenotype & anatomy of cancer) such as cellular composition, tumour density, shape & size, etc. These techniques are crucial for detection of cancer therefore there is a huge need to enhance them. With the onset of Industry 4.0, AI came into picture. As AI has a lot of potential to handle big data & to classify objects based on algorithms, therefore it can be implemented for disease diagnostics for faster results with better accuracy. Due to its overwhelming potential, AI has developed its roots in many sectors including business & market, agriculture, food, education, & even healthcare.

Cancer is not a new disease but it was always there, as identified by Hippocrates, ‘The Father of Medicine’. Cancer is the abnormal transformation of cells that exponentially grow & spread to other body parts through ‘Metastasizing’. Cancer is very deadly and is responsible for a large number of deaths globally, accounting for 10 million deaths in 2020, prominently by lung cancer (1.8 million deaths), colon & rectum cancer (916000 deaths), liver cancer (830000 deaths), stomach cancer (769000 deaths) and breast cancer (685000 deaths) [1,2], which can be seen in **Fig. 1**, respectively. Due to lack of efficient & faster diagnostic techniques, the cancer crosses the curable stages & becomes fatal. It is estimated that by 2050, there will be around 35 million new cancer cases, i.e. 77% more than 20 million cases in 2022. Due to this impending danger, urgent measures have to be taken in order to save the people from cancer. One of the approaches is AI i.e. developing at a greater rate. AI algorithms can be developed that can detect cancer at early stages which can help to fasten the treatment process.

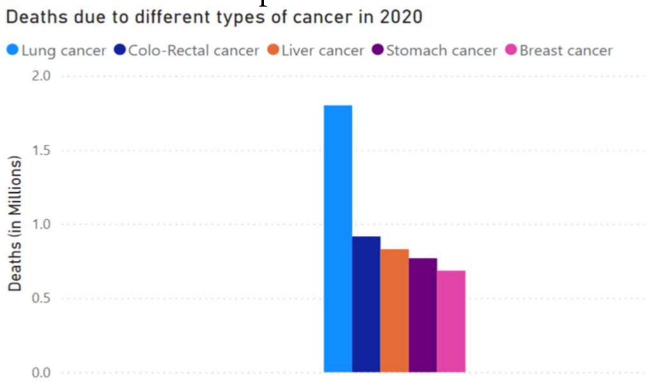


Fig. 1. Total deaths due to different types of cancer globally in 2020

As lungs cancer proves to be very fatal, therefore in this research we have utilized artificial intelligence to develop a web-based lungs cancer diagnostic system, named as “*Pulmoncare*” (“*Pulmon*” denotes lungs, “*onco*” denotes cancer and “*care*” denotes digital aid for patients) in order to assist radiologists for quick responses based on image classification.

II. LITERATURE REVIEW

The stage of cancer is very crucial for selecting the treatment of patient as different types of treatment procedures are followed for different stages of cancer. Therefore, determination of stage of cancer, i.e., staging is very crucial for treatment process. The whole diagnostic and treatment procedure of cancer is done by a team of medical specialists, i.e., known as multi-disciplinary team (MDT). MDT includes physician, radiologist, histo-pathologist, and oncologist. The workflow of cancer treatment starts from initial tests and biopsy by physicians, followed by tissue analysis by histo-pathologist and imaging techniques and analysis to determine staging by Radiologists. Once the stage is confirmed, the treatment is determined based on the stage and gets started immediately as per suggestions of oncologist. Clinical oncologists and surgeons treats initial stages of cancer through radiation therapies, and surgery while medical oncologists are specialized for higher treatment including Chemotherapy and Hormone therapy. One of the most famous procedures for determining the stage of cancer is TNM procedure, i.e., Tumour, Node, and Metastasis, which was developed by Professor Pierre Denoix in the duration of 1943-1952. TNM is based on the size and count of Tumours (zero tumour or T0, one tumour or T1, and so on), spread of cancer to local nodes and lymphatic system (zero node or N0, one node or N1, and so on), and the movement of tumour, or Metastasis (absence of metastasis or M0, and presence of metastasis or M1). Based on the TNM criteria, cancer may have four stages, namely stage 1, stage 2, stage 3 and stage 4, respectively. Stage 1 and 2 are contained within organ but initial metastasis starts from stage 3, which is fatal if not diagnosed earlier. Stage 1, 2, 3 and 4 are characterised by TNM features as (T1, N0, M0), (T2-3, N0-1, M0), (T4, N1-2, M4) and (T>4, N>4, M1), respectively. Therefore it is very crucial to determine stage of cancer at an early stage to save the life of patient. In this era, AI is being developed and implemented in cancer diagnostics to determine the staging and support the treatment efficiently [3].

Lungs cancer is one of most fatal type of cancer, which is caused due to abnormal growth of cells in lungs. This type of cancer is accelerated through smoking & environment, which harms the lungs. Most of the cases of lungs cancer are diagnosed at middle or advanced stage. Therefore, many researchers worked on using AI for lungs cancer detection. R. Bellotti et al. [4] developed a CAD system for detection of lungs cancer in CT images. The accuracy obtained was 88.5%. Van Ginnekan et al. [5] combined six CAD algorithms to detect cancer nodules. The accuracy obtained was 80%. Al-Kadi et al. [6] developed a system based on fractal texture features. The accuracy was 83.3%. Zakaria Suliman Zubi et al. [7] developed an ANN CAD system for lungs cancer detection, based on the parameters such as area, perimeter & shape of tumour. The accuracy obtained was 90%. Jinsa Kuruvilla et al. [8] developed a neural network for detection of lungs cancer, based on the statistical parameters including mean, standard deviation, skewness, kurtosis, fifth central moment & sixth central moment. The accuracy obtained with Traingdx (GD with variable learning rate & momentum) obtained was 91.1%. They have also proposed two training functions with better performance, i.e., 1st training function gives 93.3% accuracy, 100% specificity and 91.4% sensitivity, whereas, 2nd training function gives 93.3% accuracy & 0.0942 mean square error. Fatma Taher et al. [9] developed an ANN with Fuzzy clustering method to detect lungs cancer. They used high-grade segmentation & thresholding algorithms for obtaining high accuracy as gray level & contrast in images complicates the detection process. They have used 2 segmentation methods, including Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm for better segmentation. Their model proved to be a better alternative as compared to manual method of analysis of sputum samples.

III. METHODOLOGY

A. Dataset Collection & Preparation

As good quality dataset plays a huge role in deciding accuracy of the machine learning model, therefore a comparative analysis was done on various publically available datasets on open-source platforms such as *"Kaggle"*. The key factors considered while comparing datasets include dataset size, dataset quality and relevance to lung cancer diagnostics. A separate invalid dataset of random images were used as a class for training model to specifically predict irrelevant images for avoiding false predictions. The detailed information about the datasets that were selected is provided in **Table 1**, respectively.

Table 1. Datasets information

Image Category	Dataset Size	Number of Classes	Class-wise images distribution
MRI Scans	3,259 PNG Images	2 + 1 Invalid (from other source)	Cancer: 1478 Healthy: 1781 Invalid: 1250
Histopathological Images	14,925 JPEG Images	3 + 1 Invalid (from other source)	Adeno Carcinoma: 4975 Healthy: 4975 Invalid: 1250 Squamous cell carcinoma: 4975

The dataset was organized into different directories having name as their class names, in alphabetical format. The dataset was split into 80:20 ratio for training on 80% of the images and validating training on 20% test images in order to check the relevance and functioning of model on new data. Image data generator was used to increase the size & quality of dataset for generalization.

B. Machine Learning Algorithm Selection

A Convolutional Neural Networks (CNN) was trained on the datasets using transfer learning methodology where a MobileNetV2 architecture was adopted for utilizing pre-trained weights in order to enhance the accuracy. A CNN is a special type of Artificial Neural Network (ANN) which is mainly used for image & object classification due to its potential of recognizing patterns & details from an image or object. CNN can be implemented in disease diagnostics by complementing it with conventional medical imaging techniques such as radiology, MRI & CT, etc. CNN is very popular for having an optimized architecture that can be customized as per need w.r.t. the application & the use-case [10–13]. CNN Architecture includes a set of layers, primarily in the pre-defined sequence: Input Layer, alternate Convolution Layer &

Pooling Layer, Fully Connected Layer and the Output Layer, as shown in **Fig. 2**. The input layer gathers data of an image having particular dimensions & channels based on colour i.e. 3 channels for 3 primary colours including Red, Green and Blue, respectively.

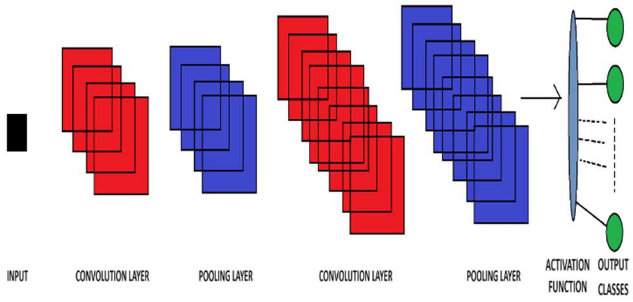


Fig. 2. Architecture of Convolutional Neural Network

The Convolution layer is the building block of whole CNN as it is responsible for doing most of the heavy computational task i.e. feature extraction. Convolution layer extracts features from an image based on pixels & channels by using matrices of different orders based on the weighted filter (mathematical function i.e. multiplied with input data to obtain desirable values, padding is done for optimizing the result). Filter is smaller in dimension as compared to input matrix, so that it's passed over different sections of input matrix to extract features.

Pooling layer is responsible for dimensionality reduction to shorten the image for better results, by reducing complexity & parameters of input through weightless filters. Pooling layer may use two common technique for dimensionality reduction, which are max-pooling & average pooling. In maximum pooling, the maximum value in input matrix is considered to be passes through the filter to form new compact matrix whereas in Average pooling, the average value of input matrix is passed into filter to form new compact matrix. Fully Connected Layer flattens the processed matrix into a vector so that it can passed into a fully connected Neural Network to obtain the desirable output based on activation function. Output layer consists the classes or categories of our desired result. Activation function returns final result in terms of numerical value such as probability, binary digits, etc. based on which the input data is classified to particular output class [14].

C. Training Machine Learning model

The key parameters considered for training the CNN models are listed in **Table 2**.

Table 2. Training parameters

S. No.	Parameter	Value
1.	Image Size	224x224 pixels resizing was used for better results
2.	Batch Size	16
3.	Loss Function	Categorical Cross Entropy
4.	Optimizer	Adam with learning rate as 1 x 10^{-4}, coupled with ReduceLROnPlateau to reduce learning rate upon constant performance
5.	Activation Function	Softmax
6.	Epochs	26 with Early Stopping and Checkpoint to avoid over training
7.	Image Data Generator	for rescaling, zooming, and flipping to achieve generalization
8.	Metrics	Accuracy and Loss curves

Optimization techniques such as learning rate reduction at constant output, early stopping for saving model upon saturation and checkpoint was implemented to enhance performance of the model and avoid overfitting. After successful training, the models were saved in .h5 format so that they can be utilized later.

D. Developing a web interface for hosting ML model

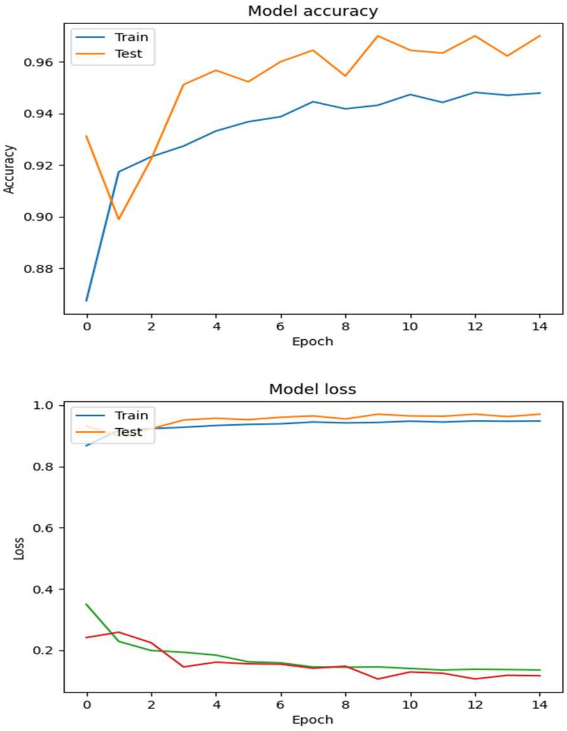
A full stack website was made using Flask as backend technology (for serving requests) and React.js as frontend technology (for providing an interface to upload image and view predicted results). Confidence threshold of 65% was implemented to filter out false predictions, and only those predictions are displayed as results which carries some relevance. Still, as this is a software trained on publically available dataset, therefore it do not claims any medication to be used without consulting actual doctor. Also, instructions regarding terms of usage were provided in order to rely only on a doctor for medication rather than trusting the results blindly.

IV. RESULTS AND DISCUSSIONS

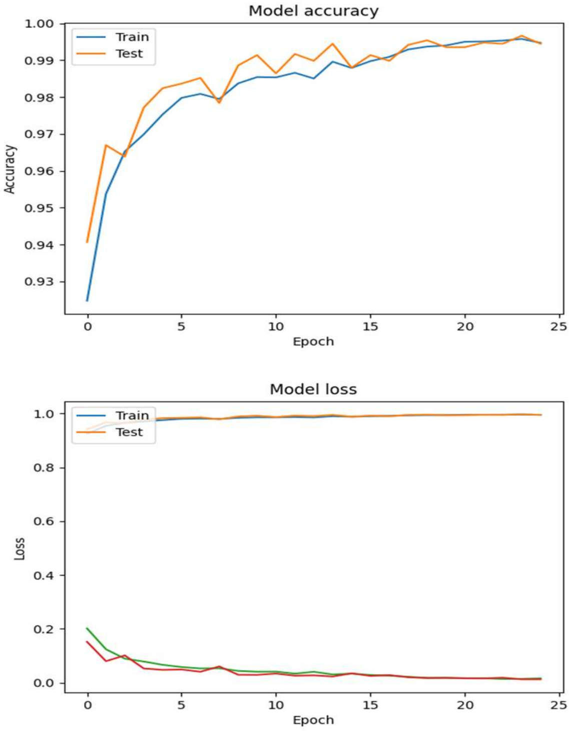
Two machine learning models were developed, one on MRI scans and other on histopathological images, with achieving validation accuracies as 96.45% and 99.57%, and validation losses as 0.12 and 0.01, as shown in **Table 3** and **Fig. 3** and **4**, respectively.

Table 3. Training results

Image Category	**Validation Accuracy**	**Validation Loss**
MRI Scans	96.45%	0.12%
Histopathological images	99.57%	0.01%

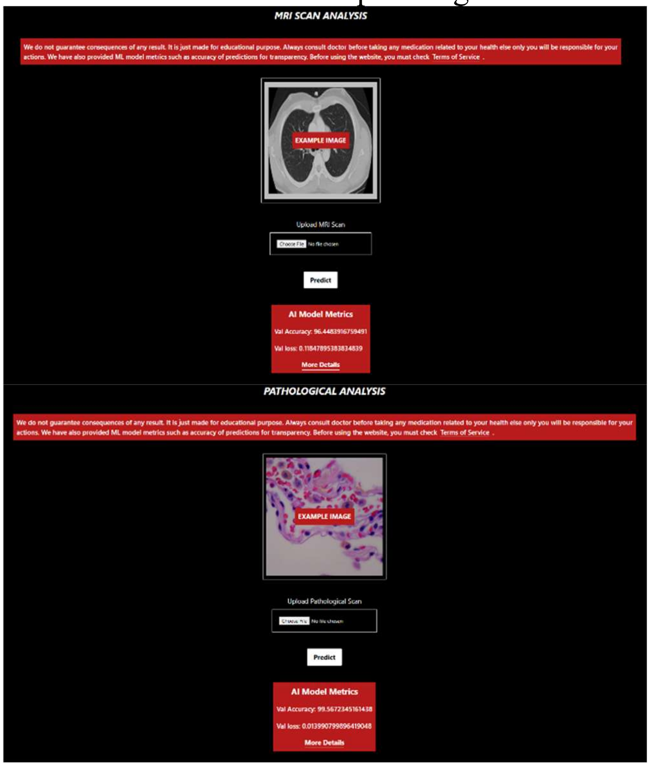


(a)



(b)

Fig. 3. Training results/metrics: (a) Accuracy & Loss curves for MRI scan's model (b) Accuracy & Loss curves for Histopathological model


MRI SCAN ANALYSIS

We do not guarantee consequences of any result. It is just made for educational purpose. Always consult doctor before taking any medication related to your health else only you will be responsible for your actions. We have also provided ML model metrics such as accuracy of predictions for transparency. Before using the website, you must check Terms of Service .

AI Model Metrics

Val Accuracy: 96.4483916759491

Val loss: 0.11847895383834839

PATHOLOGICAL ANALYSIS

We do not guarantee consequences of any result. It is just made for educational purpose. Always consult doctor before taking any medication related to your health else only you will be responsible for your actions. We have also provided ML model metrics such as accuracy of predictions for transparency. Before using the website, you must check Terms of Service .

AI Model Metrics

Val Accuracy: 99.5672345161438

Val loss: 0.013990799896419048

Fig. 4. Web interface

The deployed website is available at https://pulmoncare.vercel.app. Currently, only frontend is deployed but in later versions, we are planning to deploy backend.

V. CONCLUSION

Cancer is fatal and one of the deadliest diseases, as it's incurable at higher stages. Therefore, a need to develop faster diagnostic system for cancer diagnostics led to the implementation of AI for medical diagnostics. As AI has a great potential to manage huge data, and to develop human-like computer models based on machine learning and deep learning algorithms, therefore it is used in various sectors including medical diagnostics, digital twin technology, etc. One of the most common techniques for cancer detection is found to be Neural Networks, especially ANN and CNN. For building an optimized AI solution to any problem statement, model's training parameters, hyper-parameter tuning and optimization techniques play a crucial role.

Utilizing the knowledge from previous researches and current trends, we have developed two CNN models for lungs cancer diagnostics on MRI Scans and histopathological images, with achieving validation accuracies as 96.45% and 99.57%, and validation losses as 0.12 and 0.01, respectively. Along with that, we have developed a website named Pulmoncare to use frontend interface for using our image classification model. The novelty in this approach is that it utilizes a separate invalid images dataset as well as confidence thresholding for avoiding false predictions in order to give better results.

Based on the current understanding and statistics, AI will become more accurate in the upcoming details, which will eventually help to solve real-life problems, if yielded properly. Thereby, AI will aid early cancer diagnostics with proper and accurate results through the Computer-Aided Detection (CAD) applications.

REFERENCES

[1]. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F. Cancer statistics for the year 2020: an overview. Int J Cancer. 2021;149(4):778–789.

[2]. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7–30.

[3]. Sobin LH. TNM: evolution and relation to other prognostic factors. Semin Surg Oncol. 2003;21(1):3–7.

[4]. Bellotti R, De Carlo F, Gargano G, Tangaro S, Cascio D, Catanzariti E, et al. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. Med Phys. 2007;34(12):4901–4910.

[5]. van Ginneken B, Armato SG, de Hoop B, van Amelsvoort-van de Vorst S, Duindam T, Niemeijer M, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Med Image Anal. 2010;14(6):707–722.

[6]. Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. IEEE Trans Biomed Eng. 2008;55(7):1822–1830.

[7]. Zubi ZS, Saad RA. Improves treatment programs of lung cancer using data mining techniques. J Softw Eng Appl. 2014;7:69–77.

[8]. Kuruvilla J, Gunavathi K. Lung cancer classification using neural networks for CT images. Comput Methods Programs Biomed. 2014;113(1):202–209.

[9]. Taher F, Werghi N, Al-Ahmad H, Sammouda R. Lung cancer detection using artificial neural network and fuzzy clustering methods. Am J Biomed Eng. 2012;2(3):136–142.

[10]. Sermanet P, Chintala S, LeCun Y. Convolutional neural networks applied to house numbers digit classification. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR); 2012.

[11]. Cireşan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J. High-performance neural networks for visual object classification. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI); 2011.

[12]. Coleman S, Kerr D, Zhang Y. Image sensing and processing with convolutional neural networks. Sensors (Basel). 2022;22(10):3612.

[13]. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV); 2014.

[14]. Derry A, Krzywinski M, Altman N. Convolutional neural networks. Nat Methods. 2023;20(9):1269–1270.