Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning with Class Selective Image Processing

S.Jaipreetha[1], S.P.Mahima, S.Kaviya Linga Shree, S.Vidhya Lakshmi, Mrs.B.Bala Abirami.
Computer science and engineering Panimalar Institute of technology Chennai
M.E Assistant Professor, Computer science and engineering Panimalar Institute of technology Chennai

***Abstract*- Lung and colon Cancer could be a Disease of uncontrolled cell growth in tissues of the lung. Discovery of carcinoma in its initial stage is that the key of its cure. All in all, a measure for earlier than schedule stage lung disease determination essentially incorporates those using Histopathology images. In numerous countries collecting Histopathology images isn't yet pragmatic. So, we'll utilize some systems are key to image dataset of medicinal picture mining, Lung Field collection, processing, Feature Extraction classification utilizing transfer learning algorithm of Vgg16, the power of deep learning, to address the challenges of early- stage detection in two critical types of cancer lung and colon. By using large-scale medical imaging datasets, the VGG16 model is trainedto recognize subtle patterns.**

Our approach seeks to provide a robust and automated tool for identifying potential malignancies at their incipient stages. The fusion of cutting-edge technology with medical expertise underscores our commitment to advancing healthcare and enhancing the prospects of early cancer detection.

Manifestation Term – Feature Extraction, Transfer Learning Class-Selective Image Processing, Image Segmentation.

1. INTRODUCTION

Lung Cancer may be a noteworthy reason for Mortality within the western world as displayed by the striking factual figures distributed constantly by the American Carcinoma Society. They demonstrate that the 5- time survival rate for cases with lung malice are frequently enhanced from a standard of 14 up to 49 if the disease is analyzed and treated at its original stage. Medicinal pictures as a significant piece of remedial determination and treatment were specializing in these pictures permanently. These pictures incorporate success of concealed data that misused by doctors in opting contemplated choices around a case. This reason inspires to use information digging systems capacities for productive literacy birth & find concealed lung. Mining Medical Pictures includes numerous procedures. Medicinal data processing may be a promising zone of computational insight connected to a accordingly break down case's records going for the exposure of latest information precious for restorative choice making. This model should be able of directly relating early- stage cancer or abnormalities in lung and colon histopathology images. Enable the early discovery of cancerous or pre-cancerous lesions in lung and colon tissues. Beforehand discovery is pivotal for timely intervention and improved treatment issues. Comprehensive datasets of lung and colon histopathology images, ensuring diversity of cases.

1.2 LITERATURE SURVEY TECHNIQUES USED

1. IMAGE ACQUISITION PHASE

The first step is to acquire images. The computer need to view many images to recognize an object. Other types of data, such as time data, can also be used to train deep learning models. In the context of the work surveyed in this paper, the relevant data required to detect lung and colon cancer will be images. Images that could be used include CT scan and MRI image and these dataset collected from Kaggle. The output is images that will later used to train themodel.

2. DATA PREPROCESSING

Dynamic Histogram Equalization (DHE) to improve the quality of images before they were inputted into the CNN architectures model. Histogram Equalization (HE), which denotes mapping from the initial narrow pixel

levels to a wider extent and improves image enhancement, has widely been used in image processing. The HE technique means to convert the gray levels of an image by using cumulative effort function globally, yet always brings about the problem that elaboration information in images is damaged, leading to awful image quality.

1.1 TRANSFER LEARNING

Transfer learning is a machine learning technique where knowledge gained from solving one problem is applied to a different but related problem. In the context of neural networks, transfer learning involves using the knowledge learned by a model on one task (the source task) to improve performance on a different task (the target task).

Here's how transfer learning works:

Pre-trained Models: Transfer learning often starts with a pre-trained model that has been trained on a large dataset for a specific task, typically image classification or object detection.

Transfer of Knowledge: By leveraging the learned features and representations from the source task, transfer learning enables the model to generalize better to the target task, even when the target task has a smaller dataset or different data distribution.

3. CASSIFICATION EVALUATION METRICS

In this subsection, several evaluation metrics, accuracy, loss and so on, are described. According to the outputs of model, four indices, True Positive, True Negative, False Positive, False Negative, are used to analyze and identify the performance of model. The True Positive means that the chest CT-Scans images, which suffer from lung and colon stages, are signed as malignancy well by the model. The True Negative means if the CT images do not show malignancy level as well as the model predicts.

4. ENSEMBLE OF CLASSIFIERS

When more than one classifier is combined to make a prediction, this is known as ensemble classification. Ensemble decreases the variance of predictions, therefore making predictions that are more accurate than any individual model. From work found in the literature, the ensemble techniques used include majority voting, probability score averaging and stacking.

1.3 LITERATURE SURVEY

S.NO	OBJECTIVE	TECHNIQUES USED	DISADVANTAGES
1	Automatic feature extraction for the prediction to be accurate and to use a reliable method for that	Bi-LSTM Structure Subsidiary AttentionMechanism (S-Att) Contextual InformationExtraction	Machine learning models heavily rely on the qualityof data. Inaccurateness can lead to flawed predictions.
2	It is very easy toimplement. This gives usfriendly mode. No new hardware isneeded for this Detection.	Local Binary Patterns Bidirectional Long Short-Term Memory Deep Belief Network(DBN)	The deep learning models require large amounts of high-quality training data to achieve optimal performance.
3	To detect glaucoma at early stages with the help of deep learning- based feature extraction Retinal fundus images are utilized for the training and testing of our proposed model	CNN for feature extraction high-level feature computation from fundus images enabling DL based analysis of image pattern associated with glaucoma Local binary patterns(LBP)	The choice of algorithm and feature descriptors in hybrid approach may impact overall performance of system, selecting optimal combination requires through validation.
4	A deep neural network for detecting lung cancer from CT images is developed and evaluated	The document discuss the evaluation of proposed model using metrics such as accuracy , precision, recall & F1 score. These metrics help assess performance of deep learning model.	The use of deep learning in healthcare raises ethical & legal considerations , such as patient , privacy , data security and liability issues.
5	multi-label classification model combining attention-based neural networksand association-specific contexts is proposed forthe detection of multiplelesions on chest X-ray images	LSTM networks. CNN forfeature extraction	Requires a high computational cost. Less efficient than other algorithms. Data annotation and labelling, overfitting. Ethical considerations

6.	Those feature types that performedwell in the global phase are thenextracted from the each of these blocks, to represent each blockwith the feature vectors.	Preprocessing steps such as image warping and Cropping to standardize the inputimages	The study uses a dataset of 1000 CT scan images, access to large and diverse datasets for training and testing purposes may be a limitation for researches.
7.	To build automatic cancer prediction system that is accurate and at which stage of cancer or to improve the accuracy of the previous cancer prediction.	Thresholding procedure was used to prevent displayof unnecessary artifacts in the CTimages, ensuringthe accuracy of analysis.	Deep learning algorithms rely heavily on large amounts of high quality data for training, which may not always be readily available or easily accessible.
8.	Rare lung adeno squamous carcinoma (ASC) samples for the first time, and proposed a computer-aided diagnosis method based on the histopathological images of ASC, lung squamous cell carcinoma (LUSC) and small cell lung carcinoma (SCLC).	Adoption of relief feature selection algorithm to identify relevant features for classification	Availability of limited dataset , especially for rare lung cancer subtypes like adeno squamourrs carcinoma(ASC)
9.	The current stage of the challenge focused on lung cancer segmentation. 200 slides were used for challenge, also methods from the top 10 teamswere selected for the comparison.	Thresholding procedure was used toprevent display ofunnecessary artifacts	Subjectivity , Performance trade-oofsUser acceptance
10.	To analyze the Contribution and application offorced oscillation technique (FOT) devices in lung cancerassessment.	A Prototype 4P-FOT device operating at low-range frequencies was employed for140-second measurements toassess respiratory impedence	Reluctance of patients to participate in clinical trials and potential shortages due to pandemic-related causes can impact the recruitment and retention of participants for FOT measurements, leading to challenges in data collection and analysis
11.	Large datasets of lung and colon histopathology image was used for training, testing and validation process.	Performed lung cancer detection by employing recurrent neural network (RNN) with damped Least squares method	Inefficient algorithms, that results in higher computation cost and time utilization. Reported low accuracy. Multiple datasets with varying scan settings, resulting in the fake positive results.
12.	To increase the program prognostic assessments, eventually contributing to effective treatments.	IWSACAE-LCCD for lung and colon cancer detection. Convolutional encoder(CAE)	Creates decreased size of resultant mapping features from the output.
13.	To identify the morphological abnormalities in tissue samples of the infected area.	LC25000 lung and colon pathology image collection	It's critical to promote routine cancer screening and suggest prompt treatment with less efficient therapeutic approaches.
14.	To effectively learn and classify cancerous and non- cancerous patterns	BICLCD-TSADL applies Gabor filtering(GF) to preprocess input image	Statistical techniques is comparatively easy to model but inefficient forecast power. Needs several computations and consumes for about long-time.
15.	Effective in assisting , interpreting and analyzing lung cancer images	CNN based on ResNet-50a pre trained CNN model	For CNN to achieve cutting-edge accuracy, a large quantity of training datasets and additional time are needed.

Figure1.3 Table of literature survey

3. ARCHITECTURE

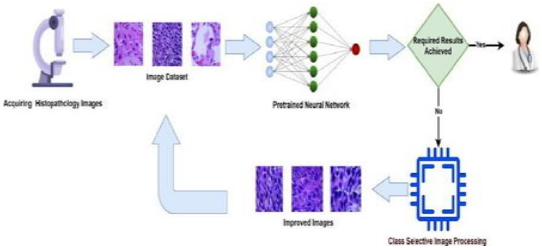

FIGURE 1: Outline classification ofhistopathology images

1. Input Histopathology Images:

These are the images of lung and colon tissues obtained through histopathology.

2. Data Preprocessing:

Preprocessing steps may include resizing, normalization, and augmentation to prepare the images for input into the model.

3. Transfer Learning:

Pre-trained models, such as those trained on large datasets like ImageNet, are utilized. This involves transferring

knowledge gained from the pre-trained model to the task of malignancy detection.

4. Feature Extraction/Fine-Tuning:

Feature extraction involves extracting relevant features from the images. Fine-tuning involves adjusting the parameters of the pre-trained model to better suit the current task.

5. Model Architecture:

This typically involves a convolutional neural network (CNN) architecture optimized for image classification tasks. It may consist of multiple convolutional layers followed by pooling layers and fullyconnected layers.

6. Image Registration/Alignment:

Register and align histopathology images from different sources or modalities to ensure consistency and accuracy in featureextraction.

7. Multi-Modal Fusion:

If available, integrate information from other modalities such as MRI or CT scans with histopathology images to improve the model's performance.

8. Uncertainty Estimation:

Implement uncertainty estimation techniques, such as Bayesian methods, to quantify the uncertainty associated with model predictions.

9. Model Ensemble/Averaging :

Combine predictions from multiple models or model snapshots to improve robustness and generalization.

10. Malignancy Detection andClassification:

Train a classifier, such as a fully connected neural network or support vector machine, on the extracted features to detect and classify malignancy in histopathology images.

Utilize appropriate loss functions and evaluation metrics (e.g., cross-entropy loss, accuracy, sensitivity, specificity) to optimize and assess model performance.

11. Output:

Generate predictions for malignancy status (malignant or benign) for each histopathology image based on the trained classifier.

12. Evaluation and Validation:

Evaluate the performance of the proposed system using held-out validation data or cross-validation techniques.

Validate the system's effectiveness and reliability through expert review and comparison with ground truthannotations.

13. Deployment and Integration:

Deploy the trained model as part of a larger system or workflow for automatedhistopathology analysis.

Integrate the system with existing medical imaging platforms or tools for seamless integration into clinical practice.

4.PROBLEM STATEMENT

Histopathology is the microscopic analysis of tissue samples to investigate the cellular manifestations of diseases. In the context of cancer detection, histopathology images often involve examining thin slices of tissue samples under a microscope to identify cancerous or pre-cancerous changes in cells.

Histopathology images are essential for diagnosing cancer and determining its stage and grade. These images can reveal important information about tissue morphology, cellular abnormalities, and the presence of cancerous cells,

By harnessing the power of deep learning, particularly through the VGG16 algorithm, this project aspires to create a transformative solution that empowers healthcare professionals with a reliable and efficient tool for early-stage cancer detection.

5.EXISTING SYSTEM

The existing system is time consuming process, and it very difficult to detect it in its early stages as its symptoms appear only within the advanced stages.

Implementing the system to automate the classification process for the first prediction of carcinoma. When compared to the after mentioned classes, the lung-SCC and lung-ACA classes performed poorly, inefficient algorithms, resulting in higher computational cost and time utilization.

Multiple datasets with various experiment settings, ensuing in false-advantageous results.

Earlier lung Disease detection studies focused on a single form of cancer, however in this study, we used our model to identify lung disease the same time.

Although few research carried out picture processing strategies to enhance the pleasant of photographs earlier than classification, however, in those research, picture processing changed into carried out to the complete dataset ensuing in higher time utilization and computing cost, etc.

6.PROPOSED SYSTEM

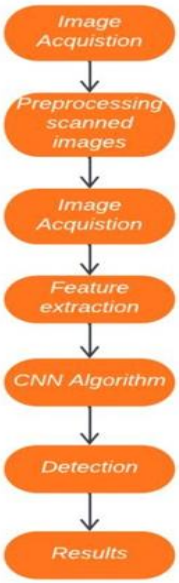

FIGURE 2: Block Diagram

In the proposed system, we address the problem of medical data scarcity by considering the task of detection of Lung and colon cancer stages from chest CT images using small volume datasets. We implemented convolutional neural networks of vgg16.

Pre-trained on the Image Net dataset and assessed them in lung disease classification tasks using transfer learning approach.

We created a pipeline that segmented chest (CT) images prior to classifying them and we compared the performance of our framework with the existingones.

We demonstrated that pre-trained models and simple classifiers such as shallow neural networks can compete with the complex systems.

Furthermore, our vgg16 based model almost tied with the best performing solution on the dataset despite being computationally less expensive.

Some of the modules used are:

- Image acquisition phase
- Data pre-processing
- Classification evaluation metrics
- Ensemble of Classifiers

IMAGE ACQUISITION PHASE:

The first step is to acquire images. To produce a classification model, the computer needs to learn by example. The computer needs to view many images to recognize an object. Other types of data, such as time data, can also be used to train deep learning models. In the context of the work surveyed in this paper, the relevant data required to detect lung and colon cancer will be images. Images that could be used include CT scan and MRI image and these datasets collected from Kaggle website. The output of this step is images that will later be used to train the model.

DATA PRE-PROCESSING:

Image pre-processing is a very common and beneficial technique in the deep learning process and it not only could enlarge the quantity of the original dataset but also enrich the information implicit in the dataset. As previously mentioned, we utilized an effective image enhancement method named Dynamic Histogram Equalization (DHE) to improve the quality of images before they were inputted into the CNN architectures model. Histogram Equalization (HE), which denotes mapping from the initial narrow pixel levels to a wider extent and improves image enhancement, has been widely used in image processing. The HE technique means to convert the gray levels of an image by using cumulative effort function globally, yet always brings about the problem that elaboration information in images is damaged, leading to awful image quality. This popular image contrast enhancement method could enhance image contrast effectively in many aspects, like CT-Scan images.

CLASSIFICATION EVALUATION METRICS:

In this subsection, several evaluation metrics, accuracy, loss and so on, are described. According to the outputs of model, four indices, True Positive, True Negative, False Positive, False Negative, are used to analyze and identify the performance of model. The True Positive means that the chest CT- Scans images, which suffer from lung and colon stages, are signed as malignancy well by the model. The True Negative means if the CT images do not show malignancy level as well as the model predicts.

The precision rate was always used to estimate howmuch the number of images that are truly cancer accounted for in the total number examples, which are classified as positive for lung and colon cancer. That is, the begin or malignant images must be identified in practical clinical diagnoses and it predicts the lung and colon stages.

ENSEMBLE OF CLASSIFIERS:

When more than one classifier is combined to make a prediction, this is known as ensemble classification. Ensemble decreases the variance of predictions, therefore making predictions that are more accurate than any individual model. From work found in the literature, the ensemble techniques used include majority voting, probability score averaging and stacking. In majority voting, every model makes a prediction for each test instance, or, in other words, votes for a class label, and the final prediction is the label that received the most votes. An alternate version of majority voting is weighted majority voting, in which the votes of certain models are deemed more important than others. For example, in probability score averaging, the prediction scores of each model are added up and divided by the number of models involved. An alternate version of this is weighted averaging, where the prediction score of each model is multiplied by the weight, and then their average is calculated. Examples of works which used probability score averaging are found. In stacking ensemble, an algorithm receives the outputs of weaker models as input and tries to learn how to best combine the input predictions to provide a better output prediction.

7.RESULT

```
model.summary()
Model: "model_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         (None, 224, 224, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168
```

FIGURE 7.1 OUTPUT

```
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
_________________________________________________________________
flatten_3 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 50178
=================================================================
Total params: 14,764,866
Trainable params: 50,178
Non-trainable params: 14,714,688
_________________________________________________________________
```

FIGURE 7.2 OUTPUT

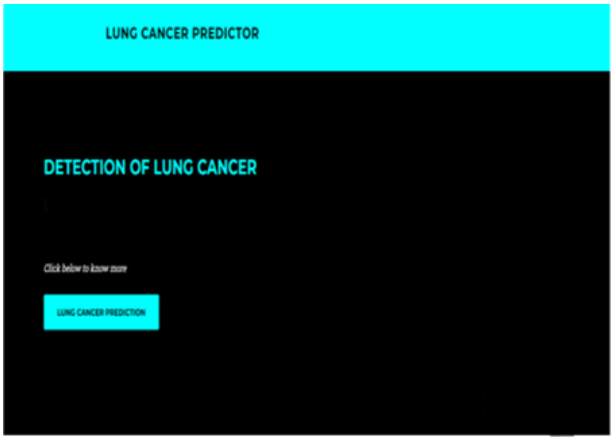

FIGURE 7.3 OUTPUT

8.CONCLUSION

Early lung cancer frequently has no symptoms and can only be detected by medical imaging. As the cancer progresses, utmost people witness nonspecific respiratory problems coughing, briefness of breath, or casket pain. Other symptoms depend on the position and size of the excrescence. Those suspected of having lung cancer generally suffer a series of imaging tests to determine the position and extent of any excrescences. Definitive opinion of lung cancer requires a vivisection of the suspected excrescence be examined by a pathologist under a microscope. In addition to feting cancerous cells, a pathologist can classify the excrescence according to the type of cells it originates from.

Around 15 of cases are small- cell lung cancer (SCLC), and the remaining 85(the non-small-cell lung cancers or NSCLC) are adenocarcinomas, scaled- cell lymphomas , and large- cell lymphomas. After opinion, farther imaging and necropsies are done to determine the cancer's stage grounded on how far it has spread. Treatment for early stage lung cancer includes surgery to remove the excrescence, occasionally followed by radiation remedy and chemotherapy to kill any remaining cancer cells. After stage cancer is treated with radiation remedy and chemotherapy alongside medicine treatments that target specific cancer subtypes. Indeed with treatment, only around 20 of people survive five times on from their opinion. Survival rates are advanced in those diagnosed at an earlier stage, diagnosed at a youngish age, and in women compared to men. utmost lung cancer cases are caused by tobacco smoking. The

remainder are caused by exposure to dangerous substances like asbestos and radon gas, or by inheritable mutations that arise by chance. Accordingly, lung cancer forestallment sweats encourage people to avoid dangerous chemicals and quit smoking. Quitting smoking both reduces one's chance of developing lung cancer and improves treatment issues in those formerly diagnosed with lung cancer. Lung cancer is the most diagnosed and deadliest cancer worldwide, with 2.2 million cases in 2020 performing in1.8 million deaths. Lung cancer is rare in those youngish than 40; the average age at opinion is 70 times, and the average age at death 72.Prevalence and issues vary extensively across the world, depending on patterns of tobacco use. Prior to the arrival of cigarette smoking in the 20th century, lung cancer was a rare complaint. In the 1950s and 1960s, adding substantiation linked lung cancer and tobacco use, climaxing in affirmations by utmost large public health bodies discouraging tobacco use. Lung cancers are among the leading causes of casualty worldwide. Beforehand and accurate opinion of these cancers can significantly ameliorate remedial issues and survival rates. The thing of this study was to descry lung and colon cancer countries of directly and efficiently. Transfer literacy is employed for this discovery of cancer on dataset CT images of lung and colon. Our stationed DL network's achieves high delicacy.

The proposed methodology has not only outperformed state- of the- art styles for lung cancer discovery in terms of delicacy but has also reduced the time and computational cost. We believe that our proposed methodology can also be intertwined in the effective opinion of other conditions.

ACKNOWLEDGEMENT

I would like to express my special thanks to our mentor Mrs.B. Bala Abirami. M.E for their time and efforts that they have provided throughout the year. Their useful advice and suggestions were really helpful to us during the project's completion. In this aspect, we are eternally grateful to them. I would like to acknowledge that this project was completedentirely by me and not by someone else.

REFERENCE

[1] S. SAbbas and M. M. Q. A. Yasin, ''Lungs cancer detection using convolutional neural network,'' Int. J. Recent Adv. Multidisciplinary Topics, vol. 3, no. 4, pp. 90–92, 2022.

[2] A. Khan, ''Identification of lung cancer using convolutional neural networks based classification,'' Turkish J. Comput. MathS Educ. (TURCOMAT), vol. 12, no. 10, pp. 192–203, 2021

[3] C. Thallam, A. Peruboyina. T. Raju, and N. Sampath, ''Early stage lung cancer prediction using various machine learning techniques,'' in Proc. 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Nov. 2020, pp. 1285–1292. [4]Y.Xie,W.-Y.Meng,R_Z.Li,YW. Wang,X.Qain,C.Chan,Z.F.Yu,X.-X.Fan,H.- D.Pan,C.Xie,Q-B.Wu,P-y.yan,L.-Liu,Tang,Yao,M-F Wang,and E.L.H. Leung,"Early Lung cancer diagnostic biomarker discover by machine learning methods, translational oncol.,vol.14, no.1,Jan.2021,Art NO.100907.

[5] R.Mahum, S. U. Rehman, O. D. Okon, A. Alabrah, T. Meraj, and H. T. Rauf, ''A novel hybrid approach based on deep CNN to detect glaucoma using fundus imaging,'' Electronics, vol. 11, no. 1, p. 26, Dec. 2021.

[6] N. Kalaivani, ''Deep learning based lung cancer detection and classification,''IOP Conf.Ser.,Mater. Sci. Eng., vol. 994, no. 1, 2020, Art. no. 12026

[8] L. Ale, N. Zhang, and L. Li, ''Road damage detection using RetinaNet,''in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2018, pp. 5197–5200.

[9] J.Redmon and A.Farhadi,''YOLOv3:An incremental improvement,''2018,arXiv:180402767.

[10] J. Chaki and M. Woźniak, ''Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review,'' Biomed. Signal Process. Control, vol. 80, Feb. 2023, Art. no. 104223.