LUNG CANCER DETECTION USING IMAGE PROCESSING AND DEEP LEARNING METHODS

By

JIYA JIVEESHA	**21BCE1255**
MEELA AKSHAYA	**21BCE1987**

A project report submitted to

Professor Dr. Rajalakshmi

BCSE209L – MACHINE LEARNING

School of Computer Science and Engineering (SCOPE)

TABLE OF CONTENTS

ABSTRACT

This project aims to create a model using deep learning that can detect lung cancer at an earlier stage. A Convolutional Neural Network architecture is used to analyse the medical images of the lungs to classify them as malignant or benign. The dataset used in the study comprises CT scan images of lung nodules from the publicly accessible LIDC-IDRI dataset. Transfer learning is used with the previously taught VGG-16 architecture to train the CNN model. The accuracy, precision, recall, and F1 score are some of the standard metrics utilised to evaluate the suggested model's performance. The findings show that the proposed model is capable of high accuracy and outperforms the existing methods that are considered to be state-of-the-art. The model that has been proposed has the potential to be of assistance to radiologists in the process of the early detection of lung cancer and to enhance the results for patients.

Keywords: Image processing, deep learning, lung cancer, CNN, performance metrics

INTRODUCTION

Lung cancer is a form of cancer that develops in the lung tissues, typically in the cells lining the airways. It occurs when abnormal cells in one or both lungs develop uncontrollably and can spread to other organs. Lung cancer is one of the most prevalent types of cancer and the largest cause of cancer-related deaths around the world. It is a severe condition that can be challenging to treat, particularly if it is not discovered in its early stages. Lung cancer is frequently identified at an advanced stage, making it difficult to treat and limiting the likelihood of survival. There are two primary forms of lung cancer, the most prevalent of which is non–small cell lung cancer (NSCLC). Small cell lung cancer (SCLC) accounts for around 20% of all lung cancers. If both types are present, the lung cancer is referred to as small mixed cell/large cell cancer. If cancer originates elsewhere in the body and travels to the lungs, this is known as metastatic lung cancer.

Cigarette smoking is the leading cause of lung cancer, accounting for around 85 percent of cases. Cigarette smoke contains over 70 recognised carcinogens, such as polycyclic aromatic hydrocarbons (PAHs), nitrosamines, and benzene, which can damage lung DNA and lead to cancer. Exposure to secondhand smoking can also raise the chance of developing lung cancer, especially in nonsmokers exposed to high levels of secondhand smoke for an extended length of time. Environmental contaminants, such as radon, asbestos, and diesel exhaust, can also damage the DNA in lung cells and raise the risk of cancer. In extremely rare instances, lung cancer can be caused by inherited genetic abnormalities. Age, gender, and family history can also increase the likelihood of acquiring lung cancer. Limiting exposure to established risk factors, such as stopping smoking and avoiding environmental contaminants, can help reduce the incidence of lung cancer. Lung cancer is a relatively contemporary illness. Unfortunately, the illness did not gain extensive notice and research until the 20th century.

In the 1930s, researchers initially discovered a greater incidence of lung cancer in smokers compared to nonsmokers, establishing the connection between smoking and lung cancer. This revelation prompted a public health initiative to discourage smoking and reduce lung cancer incidence. In the 1950s and 1960s, breakthroughs in medical technology led to the creation of X-ray machines and other imaging instruments that might identify lung cancer. This resulted in an increase in early detection and an improvement in survival rates. In the 1970s, researchers began classifying lung cancer subtypes based on their microscopic

appearance, resulting in more specialized therapy options. Throughout the 1990s and 2000s, novel medications and therapies, including targeted therapies and immunotherapies, were developed to treat lung cancer. Detecting lung cancer early is essential for improving outcomes and boosting survival chances. Imaging techniques such as chest X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI) are used to diagnose lung cancer (MRI). In addition, persons with a high risk of getting lung cancer, such as current or past smokers, are advised to undergo screenings such as low-dose CT scans. Biomarker testing, which involves the analysis of specific proteins or genetic markers in a patient's blood, tissue, or other bodily fluids, is an additional promising method for the early diagnosis of lung cancer.

Computer-aided systems can play a crucial role in the detection of lung cancer by supporting healthcare practitioners in recognising suspicious lesions or nodules in medical pictures of the lungs. Early lung cancer detection can increase the likelihood of successful therapy and lower mortality risk. Computer-aided systems can aid in the detection of small lung nodules or lesions that may be difficult to identify with conventional radiography.

Computer aided systems can accurately evaluate images of the lungs, lowering the likelihood of human mistake during the detection procedure. These systems employ machine learning algorithms capable of analysing large volumes of data and identifying patterns that the human eye might overlook. Conventional lung cancer screening procedures can be time-consuming, requiring numerous imaging and diagnosis rounds. By rapidly recognising worrisome nodules or lesions, computer-aided systems can reduce the time required to diagnose and treat lung cancer. Employing computer-aided systems for lung cancer identification can lower the costs associated with conventional diagnostic procedures, such as recurrent imaging and biopsy procedures. CAD systems can also aid in the interpretation of medical pictures by radiologists. These tools can help radiologists make more accurate and quick diagnosis by giving more information and indicating potentially worrisome areas. In addition, computer-aided diagnosis (CADx) systems can analyse additional patient data to generate a probability score for the presence or absence of cancer. This can help direct clinical decision-making and decrease the amount of needless biopsies and invasive treatments.

Researchers are aiming to increase the accuracy and efficiency of lung cancer detection methods through the development of new technologies and methodologies, which will

ultimately save lives and improve patient outcomes. The application of machine learning, a subset of artificial intelligence, to improve the accuracy and efficiency of lung cancer diagnosis has shown considerable promise. Machine learning (ML) is a promising method for detecting lung cancer that has received much attention in recent years. Large datasets of patient information, including imaging data and biomarker measurements, can be used to train machine learning algorithms in order to construct models that effectively predict the existence of lung cancer.

With imaging data related with lung cancer, machine learning algorithms can also recognise certain patterns, such as the size, shape, and texture of tumors. In addition, machine learning algorithms can scan vast quantities of genetic data to uncover potential lung cancer indicators. These biomarkers can be utilised to build diagnostic assays for the early detection of lung cancer. Deep learning (DL) is a fast-developing discipline of artificial intelligence that has demonstrated promise in detecting and diagnosing lung cancer. Deep learning (DL) is a subfield of machine learning that has the potential to improve lung cancer detection's precision and efficacy. DL algorithms employ artificial neural networks that are capable of learning from vast volumes of data and identifying patterns that are difficult to spot using conventional approaches.

The use of DL to detect lung cancer from radiological images such as CT scans has shown remarkable promise. CNNs, a type of DL algorithm, can automatically recognise and segment lung nodules, which are frequently the earliest sign of lung cancer. In addition, DL can increase the precision of discriminating benign from malignant nodules and forecast the chance of cancer metastasis.

The rest of this is paper organized as follows: Literature Review presents the research work related to the domain of detecting lung cancer by image analysis and other methods. Proposed Architecture elaborates the proposed approach for classifying lung cancer along with other classes. It also includes data set description which describes the data set used in the study to classify various classes lung cancer. Results section shows figures that are processed during project and interpretation of them. Results analysis discusses the implementation of various parts of the proposed architecture followed by Conclusion.

RELATED WORKS

Image processing based detection of lung cancer on CT scan images

In this paper, they implemented and analyzed the image processing method for detection of lung cancer. Image processing techniques are widely used in several medical problems for picture enhancement in the detection phase to support the early medical treatment. In this research they proposed a detection method of lung cancer based on image segmentation. Image segmentation is one of intermediate level in image processing. Marker control watershed and region growing approach are used to segment of CT scan image. Detection phases are followed by image enhancement using Gabor filter, image segmentation, and features extraction. From the experimental results, they found the effectiveness of our approach. The results show that the best approach for main features detection is watershed with masking method which has high accuracy and robust.

Lung Cancer Detection using CT Scan Images

Lung cancer is one of the dangerous and life taking disease in the world. However, early diagnosis and treatment can save life. Although, CT scan imaging is best imaging technique in medical field, it is difficult for doctors to interpret and identify the cancer from CT scan images. Therefore computer aided diagnosis can be helpful for doctors to identify the cancerous cells accurately. Many computer aided techniques using image processing and machine learning has been researched and implemented. The main aim of this research is to evaluate the various computer-aided techniques, analyzing the current best technique and finding out their limitation and drawbacks and finally proposing the new model with improvements in the current best model. The method used was that lung cancer detection techniques were sorted and listed on the basis of their detection accuracy. The techniques were analyzed on each step and overall limitation, drawbacks were pointed out. It is found that some has low accuracy and some has higher accuracy but not nearer to 100%. Therefore, our research targets to increase the accuracy towards 100%.

LITERATURE REVIEW

Rahen et al [1] created a lung cancer detection system in order to classify the existence of lung cancer in CT scans and blood samples, by combining image processing and machine learning. In the study, classification is done using image feature extraction. By employing SVM and image processing techniques, a successful method for identifying lung cancer and its phases is provided, and the research also provides an overview of current systems. Some of the techniques used to develop the work model used in the study are: Image acquisition, Grayscale conversion, Noise Reduction, Binarization, Image Segmentation, Feature Extraction. In the proposed system, the user's initial CT scan images are obtained. Since CT scan replicas have less noise than X-ray and WIRI pictures, they are employed as input in order to obtain better accuracy and less distortion. The fundamental goal of segmentation is to make the delegation of a CT scan replica simple, change it into something more informative, and make it easier to analyse it in detail. The supplied CT scan image is pre-processed using each of these techniques. The main CT scan image is used to define ROI. The median filter and segmentation provide accurate results for the pre-processing phases. Some features, such as Area, Perimeter, and Eccentricity, are extracted from the ROI. These characteristics, according to study, can be used to identify lung cancer in its earliest stages. The positive and negative samples of lung cancer imaging samples are classified in this system using a Support Vector Machine classifier for grouping purposes.

Siebert et al [2] created and evaluated a model for predicting lung tumor response to IGRT therapy using sequential megavoltage CT (MVCT). For this study, tumor responses to helical tomo treatment at doses ranging from 2.0 to 2.5 Gy per fraction were assessed for 20 lung cancer lesions in 17 individuals. One patient received preoperative chemotherapy, while five patients received concurrent treatment. 480 serial MVCT images that were taken prior to treatment were contoured in order to quantify the tumor response to treatment retrospectively. Based on scant data made during the first two weeks of treatment, a regression model is constructed to predict future tumor volumes and the corresponding confidence intervals. In the proposed system, 480 serial MVCT images that were

taken prior to treatment were contoured in order to quantify the tumor response to treatment retrospectively. Before the 17 patients received 480 lung therapy fractions, megavoltage CT scans were taken. The MVCT pictures' main objective was to position the patients for treatment utilising automatic CT-to-CT fusion with the CT images used for treatment planning (20). According to the study, 5 (23%) of the 22 lung cancer lesions that were available were removed from consideration because they largely had mediastinal disease that was invisible on MVCT imaging. All forecasts were accompanied by 95% confidence intervals (CIs), which were calculated analytically. The information produced by the model can ultimately be used to make decisions about the patient's treatment because the predictability of the forecast is known.

Raoof et al [3] focused on improving the progression and treatment of malignant illnesses using machine learning techniques because of their accuracy. The current application of ML algorithms and factors that contribute to lung cancer are examined in this study, along with their relative advantages and disadvantages. In order to analyse and predict lung cancer, the healthcare industry has used a variety of machine learning techniques, including Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Network (ANN). Instead of referring to a large number of publications, this research aids the researchers in quickly reviewing the pertinent literature. In this study, a survey on lung cancer, its causes, symptoms, and cancer mortality rates in India and throughout the world is undertaken, and machine learning techniques, their applications in healthcare, as well as cancer prognosis and diagnosis, are discussed.

Alam et al [4] suggested a multi-class SVM (Support Vector Machine) classifier based approach for the accurate diagnosis and prediction of lung cancer. The classification of cancer was done in multiple stages. Lung cancer risk can also be predicted by this technique. Image enhancement and segmentation have been carried out independently at each classification stage. Image enhancement techniques include image scaling, colour space conversion, and contrast enhancement. Segmentation has been done using a watershed-based approach controlled by thresholds and markers. SVM binary classifier was employed for classification purposes. The research claims that the suggested method is more accurate at detecting and forecasting lung cancer. According to the research, the

algorithm is designed to be able to determine whether or not the input material contains tumour cells and is prepared to foresee any potential growth. The accuracy of the suggested approach is 97% for cancer detection and 87% for cancer prediction. The suggested system is effective in assisting the physician in identifying if the lung is dangerous or non-carcinogenic. By setting up the system on a large set of images and arranging it in light of genetic algorithm hereditary computation and deep neural network, the system's precision can be extemporised.

Luo et al [5] created prognostic prediction models for patients with lung cancer based on morphological features. In this study, the morphological characteristics of abnormal pictures for patients with NSCLC are examined using objective and quantitative computational methodologies. From The Cancer Genome Atlas lung cancer cohorts, tissue pathology pictures for 523 patients with adenocarcinoma (ADC) and 511 individuals with squamous cell carcinoma (SCC) were studied. In order to create statistical models that predict patient survival outcomes in ADC and SCC, respectively, characteristics from the pathological pictures were retrieved. This study is a pioneering investigation into the viability of applying digital pathological image processing for unbiased and objective clinical prognosis of lung cancer patients. The prognosis for patients with SCC and ADC is predicted by image feature-based prediction models. The research paper's suggested and used approach, meanwhile, still has certain short-term limits and offers intriguing options for future study and application.

Nageswaran et al [6] used machine learning and image processing technology to demonstrate the precise categorization and prediction of lung cancer. The correct classification of photos of illnesses is greatly influenced by the image preprocessing. Images from CT scans were with a wide range of aberrations, including noise. Image filtering techniques are used to get rid of these artefacts. To reduce the amount of noise, a geometric mean filter is applied to the input images. The space needed for the initial data matrix is reduced by applying linear discriminant analysis (LDA), which is how this is performed. Examples of parallel transformation methods include the PCA and LDA. The PCA is an unsupervised analysis technique, as opposed to the supervised LDA technique. The study showed that lung cancer is more accurately predicted by ANN. Systems for

detecting lung cancer that use robust categorization and prediction algorithms may benefit from the increased accuracy provided by this research. For implementation, this study provides cutting-edge images created using machine learning techniques.

Ausawalaithong et al [7] investigated the use of transfer learning and the 121-layer convolutional neural network, commonly known as DenseNet121, to classify lung cancer from chest x-ray pictures. To get over the issue of utilising a limited dataset, the model was trained on a dataset of lung nodules before training on the dataset of lung cancer. The suggested model also offers a heatmap to show where the lung nodule is located. These results show promise for further deep learning-based lung cancer diagnosis research using chest x-rays. The performance of the model in this study was evaluated using accuracy, specificity, and sensitivity. The performance of Model A on classifying lung nodules was evaluated for accuracy, specificity and sensitivity using a test set of the ChestX-ray14 Dataset. The threshold of Model A is 0.55 as it gives highest value between specificity and sensitivity. The performance of Models B and C on classifying lung cancer were likewise evaluated applying average and standard deviations of accuracy, specificity and sensitivity using a test set of the JSRT dataset for 10-fold cross-validation. The threshold of both models was noted as 0.5. Model B showed higher specificity but poorer accuracy and sensitivity than Model C. In addition, Model C also had lower standard deviation in all evaluation metrics. The findings show that retraining the model several times for specific tasks gives better results in almost all metrics.

Hawkins et al [8] used radiomics to choose 3-D characteristics from lung CT images in order to provide predictive data. It is demonstrated that classifiers can be created to predict survival time by focusing on instances of the adenocarcinoma non-small cell lung cancer tumor subtype from a wider data set. This is the first known study to use CT images of lung cancer to generate such predictions. Finally, they contrast feature selection methodologies and classifiers. To create a 3D segmented tumour object, the tumour is identified and segmented. Following feature extraction, classification results are obtained by applying prediction models. De-identified CT scan images from the Moffitt Cancer Center in Tampa made up the data collection that was utilized. Digital Imaging and Communications in Medicine, or DICOM, is the format used by the images.

Patients with the tumor types adenocarcinoma and squamous-cell carcinoma make up the data collection. Patients with adenocarcinoma were the subject of the paper. For the survival time study, 81 individuals with adenocarcinoma had CT scans. Using the top 5 features discovered by Relief-f, the decision tree classifier produced the maximum classification accuracy of 77.5%. The decision trees with 10 features, selected by Relief-f, have the highest AUC (0.732). There were frequently few spots on the curve for both rule learners and decision trees. Every feature is chosen per fold. CFS occasionally failed when choosing test-retest features, so those findings were left out.

Senthil et al [9] implemented a computer-aided classification technique for lung cancer prediction based on an evolutionary system using a combination of architectural evolution, weight learning using neural networks, and particle swarm optimization, this paper. This approach offered various variations and hybridized it with evolutionary algorithm to increase its performance. It employs local searching capability of neural network and global searching of PSO delivers superior lung cancer prediction as cancerous and non-cancerous. The classification process was carried out, and the outcomes were assessed using performance evaluations of several methods. The doctors can make an informed judgement depending on the patient's condition with the help of this prediction method. The performance evaluation of proposed Neural Networks with Particle Swarm Optimization are simulated using MATLAB under windows environment. The proposed method is effectively compared with K-Nearest Neighbor Support Vector Machine, Bayes Network and Neural Network in terms of performance metrics. The proposed method achieves Classification accuracy of 97.5% and it predict the accurate cancerous cells. The proposed method achieves effective results when compared to other classifiers.

Shanthi et al [10] Used image processing techniques, lung scan histopathology images to categorise lung cancer. The technique uses the features that are retrieved from lung scans for prediction. This study makes use of the grey level co-occurrence matrix and feature extraction techniques for the Gabor filter. The modified stochastic diffusion search (SDS) technique is used in this study to offer a unique wrapper-based feature selection algorithm. For classification, the decision tree, the neural network, and Naive Bayes have all been employed. In the study

140 normal and 130 abnormal images were used from the cancer genome atlas (TCGA) dataset, in this investigation. For symbolic data state, histology of lungs were obtained over maximum of 3 different time intervals. The findings demonstrate that an SDS-NN had classification accuracy of about 2.51% for an SDS-decision tree and further by about 1.25% for an SDS-Naive Bayes. It is evident that feature selection enhances the classification of images; nonetheless, more research is required to refine the classifier. This study concentrated on feature selection, however additional research into pre-processing techniques like noise removal and best classifiers is possible.

Suresh et al [11] study aims to automatically extract the self-learned features using an end-to-end learning CNN and compares the results with the conventional state-of-art and traditional computer-aided diagnosis system's performance. For the input layer, lung nodule CT images are acquired from the Lung Image Database Consortium public repository having 1018 cases. Images were pre-processed to uniquely segment the nodule region of interest (NROI) in correspondence to four radiologists' annotations and markings describing the coordinates and ground-truth values. A two-dimensional set of resampled images of size 52 by 52 pixels with random translation, rotation, and scaling corresponding to the NROI were generated as input samples. In addition, generative adversarial networks (GANs) are employed to generate additional images with similar characteristics as pulmonary nodules. CNNs was trained using images generated by GAN and are fine-tuned with actual input samples to differentiate and classify the lung nodules based on the classification strategy. In the study proposed CNN achieved the classification accuracy of 93.9%, an average specificity of 93%, and an average sensitivity of 93.4% with reduced false positives and evaluated the area under the receiver operating characteristic curve with the highest observed value of 0.934 using the GAN generated images.

Wang et al [12] aimed to provide an overview of current and potential applications for AI methods in pathology image analysis, with an emphasis on lung cancer. The study outlined the current challenges and opportunities in lung cancer pathology image analysis, discussed the recent deep learning developments that could potentially impact digital pathology in lung cancer, and summarized the existing applications of deep learning algorithms in lung cancer diagnosis and prognosis.

The study showed that with the advance of technology, digital pathology could have great potential impacts in lung cancer patient care. In the study some promising future directions for lung cancer pathology image analysis, including multi-task learning, transfer learning, and model interpretation.

Coudray et al [13] trained a deep convolutional neural network (inception v3) on whole-slide images obtained from The Cancer Genome Atlas to accurately and automatically classify them into LUAD, LUSC or normal lung tissue. The performance of the method employed was comparable to that of pathologists, with an average area under the curve (AUC) of 0.97. The model was validated on independent datasets of frozen tissues, formalin-fixed paraffin-embedded tissues and biopsies. Furthermore, it was trained the network to predict the ten most commonly mutated genes in LUAD. In the study it was found that six of them—STK11, EGFR, FAT1, SETBP1, KRAS and TP53—can be predicted from pathology images, with AUCs from 0.733 to 0.856 as measured on a held-out population. These findings suggest that deep-learning models can assist pathologists in the detection of cancer subtype or gene mutations.

Hussain et al [14] aimed to improve lung cancer image quality by utilizing and employing various image enhancement methods, such as image adjustment, gamma correction, contrast stretching, thresholding, and histogram equalization methods. We extracted the gray-level co-occurrence matrix (GLCM) features on enhancement images, and applied and optimized vigorous machine learning classification algorithms, such as the decision tree (DT), naïve Bayes, support vector machine (SVM) with Gaussian, radial base function (RBF), and polynomial. Without the image enhancement method, the highest performance was obtained using SVM, polynomial, and RBF, with accuracy of (99.89%). The image enhancement methods, such as image adjustment, contrast stretching at threshold (0.02, 0.98), and gamma correction at gamma value of 0.9, improved the prediction performance of our analysis on 945 images provided by the Lung Cancer Alliance MRI dataset, which yielded 100% accuracy and 1.00 of AUC using SVM, RBF, and polynomial kernels. The results revealed that the proposed methodology can be very helpful to improve the lung cancer prediction for further diagnosis and prognosis by expert radiologists to decrease the mortality rate.

Yu et al [15] obtained 2,186 haematoxylin and eosin stained histopathology whole-slide images of lung adenocarcinoma and squamous cell carcinoma patients from The Cancer Genome Atlas (TCGA), and 294 additional images from Stanford Tissue Microarray (TMA) Database. They extracted 9,879 quantitative image features and use regularized machine-learning methods to select the top features and to distinguish shorter-term survivors from longer-term survivors with stage I adenocarcinoma (Po0.003) or squamous cell carcinoma (P¼0.023) in the TCGA data set. The study also validated the survival prediction framework with the TMA cohort (Po0.036 for both tumour types). Results of the study suggests that automatically derived image features can predict the prognosis of lung cancer patients and thereby contribute to precision oncology. Also it was concluded that the methods employed during the research are also extensible to histopathology images of other organs.

PROPOSED METHODOLOGY

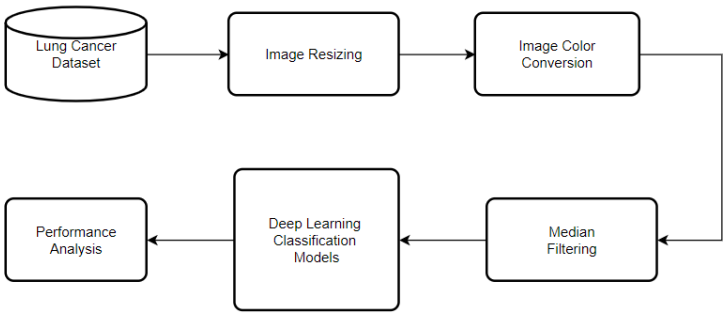

Fig 1: Proposed Architecture for Lung Cancer Detection

The Proposed Architecture has been divided into 4 main parts: image acquisition, image preprocessing, model building, and performance analysis. The first step, image acquisition, involves capturing the raw image data using a camera or sensor. The next step, image preprocessing, involves applying various techniques such as image resizing, resizing to prepare the image for analysis. The third step, model building, involves developing a deep learning or transfer learning model that can learn from the preprocessed images and perform the desired task of lung cancer classification. Finally, the performance analysis step involves evaluating the performance of the model by measuring metrics such as accuracy, precision, recall, and F1-score.

Dataset Collection:

The dataset utilised in the study is the "Lung and Colon Cancer Histopathological Images" dataset [16], which is accessible on Kaggle. It includes histopathological images of lung and colon cancer samples. Andrew A. Borkowski generated the dataset by compiling and labelling the images. The collection consists of 250,000 histopathological images classified into five classifications. The images have a 768 × 768 pixels resolution and are saved in JPEG format. Each class in the dataset consists of 5000 instances. The five classes are lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue. The images are identified with their respective cancer and tissue

types, grouped as benign or malignant. The images of lung cancer are classified according to the five classes.

The images were generated from an original sample of HIPAA-compliant and validated sources, which included 750 images of lung tissue with 250 benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas, and 500 images of colon tissue with 250 benign colon tissue and 250 colon adenocarcinomas; these were augmented to 25,000 using an augmentor package.

Image Processing:

Changing an image's size is a typical image processing technique known as image Resizing. Before feeding input images into a neural network in the context of deep learning, image resizing is frequently employed to standardize the size of the images. Resizing ensures that all input images have the same dimensions because the majority of neural networks have a fixed input size.

The cvtColor function in OpenCV is a method for converting images from one color space to another. It can be used to change the color format of an image, in this project it is used to convert BGR to RGB. It is an essential tool for image processing tasks that involve color space manipulation. It allows for the creation of images in various color spaces and enables the application of specific color transformations to images.

Reshaping is an important technique in image processing as it helps to transform the dimensionality of an image while preserving the information content. This technique can be used to transform images from one shape to another, depending on the specific needs of the image processing application. For example, reshaping can be used to convert an image from a one-dimensional array to a two-dimensional matrix, which is more suitable for convolutional neural networks (CNNs). Overall, reshaping is a versatile technique that can be applied in many different image processing applications to improve the performance and accuracy of the algorithms used.

The train-test split is a technique used in image processing to divide the dataset into two subsets: one for training the model and another for evaluating its performance. The goal of this split is to avoid the model over fitting on the training data, which is when the model becomes overly specialized to the training data and struggles to generalize to new data. The model's performance can be assessed on fresh, untested data by dividing the data into two subsets, allowing the model to be trained on one subset and tested on the other.

Additionally, the train-test split enables model selection and hyper parameter tuning, where several models and parameters may be contrasted based on how well they perform on the test set. Ultimately, the train-test split is an essential step in developing accurate and reliable models for image processing tasks.

Model Building:

CNN

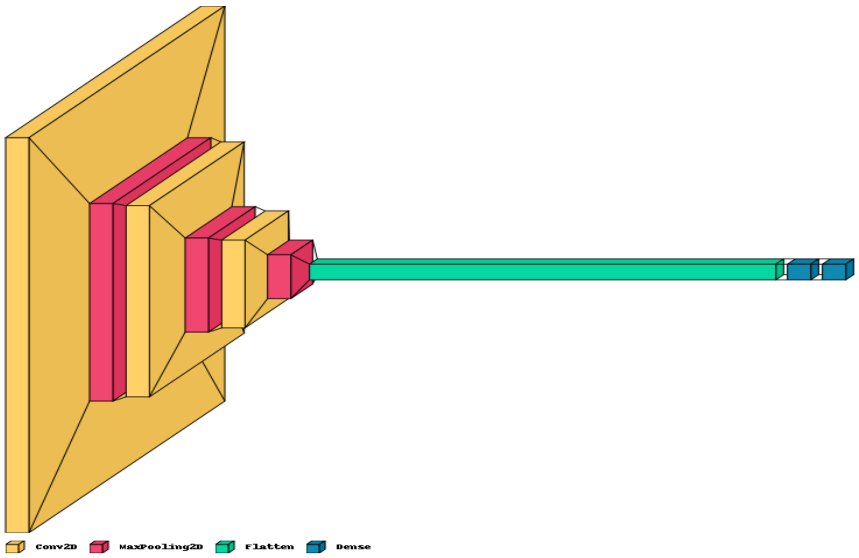

Fig 2: CNN Architecture

In order to interpret and categories visual images, neural networks of a certain type called convolutional neural networks (CNN) are used. In image recognition tasks including object detection, segmentation, and classification, it is frequently utilized. In this experiment, the lung cancer detection model was trained using a

model as the basic model. The Convolutional Neural Network (CNN) used in the lung cancer detection project involves several layers. The first layer is the input layer, which accepts the input image data. The input layer is followed by a series of convolutional layers, each of which performs a convolution operation on the input data to produce a set of output features.

The output of the convolutional layers is then sent into a pooling layer, which takes the maximum or average value of a region of the feature map to down sample the output features. The output features are flattened into a one-dimensional vector after the pooling layer and transmitted through a fully connected layer, which is in charge of understanding the more complex characteristics of the data.

The fully connected layer's output is then run through a softmax function, which normalises it. The CNN is trained with stochastic gradient descent (SGD), a variation of backpropagation that optimises the network weights and reduces the error between the expected and actual output labels.

Efficient B0

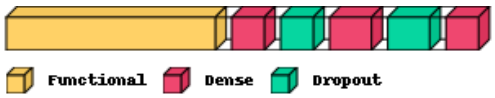

Fig 3: Efficient B0 Architecture

A deep neural network architecture called EfficientB0 is developed to improve accuracy while utilizing fewer parameters and processing resources. To lower the computational cost while maintaining high accuracy, it combines bottleneck layers, squeeze-and-excitation modules, and depth-wise separable convolutions. In order to fine-tune the model for lung cancer diagnosis in this research, EfficientB0 was employed as a pre-trained model for transfer learning.

EfficientNetB0's architecture is made up of seven convolutional layer blocks. A stem convolutional layer, the first building block, uses the input image to extract basic features. The remaining six blocks are composed of a downsampling layer, a succession of convolutional layers, and the stem layer. The squeeze-and-excitation

(SE) blocks used by EfficientNetB0 also learn to amplify relevant feature maps while suppressing less significant ones. The final classification output is generated by a fully connected layer after that. In comparison to other cutting-edge models, EfficientNetB0 has less parameters overall yet still performs very accurate on a variety of image classification tasks.

Performance Analysis:

Accuracy

The number of accurate predictions made by the model divided by the total number of predictions is used to calculate accuracy. A higher accuracy score means the model is producing more accurate predictions, whereas a lower accuracy score means the model is producing more inaccurate ones.

Accuracy might not be enough to assess a model's performance in its whole. For instance, a model that consistently predicts samples from one class in a severely unbalanced dataset might get a high accuracy score but be useless in real-world applications.

Precision

It is a performance indicator that assesses the ratio of true positives to the total of true positives and false positives. An accurate identification of instances of the target class by the model is indicated by a high precision score, whereas a high proportion of false positive predictions is indicated by a low precision score.

In cases where false positives can be very expensive, like in medical diagnosis or fraud detection, precision is extremely helpful. Even though it might miss some actual positives in these circumstances, a high accuracy model is desirable.

Recall

Recall is measured as the ratio of true positives to the total of true positives and false negatives, to be more precise. A high recall score shows that the model is correctly identifying a considerable portion of target class instances, whereas a low recall score shows that the model is missing a sizable portion of the target class instances.

When false negatives are highly expensive, as they are in medical diagnostics or security applications, recall is especially helpful. Even though it might produce more false positive predictions in certain circumstances, a high recall model is desirable.

F1-Score

A binary classification model, which predicts two classes, often indicated as positive and negative, uses the F1 score as a performance indicator to assess its overall effectiveness. The F1 score is determined as the weighted average of precision and recall.

A high F1 score indicates that the model is performing well in terms of both precision and recall, and it is useful for situations where both precision and recall are important, such as in medical diagnoses or fraud detection.

COMPARATIVE STUDY

Image processing based detection of lung cancer on CT scan images

Summary-

- **Region Growing Segmentation:**

- Select target area (right and left lung) and place seed.
- Enlarge seed to cover desired areas.
- Figure 6 shows segmentation results.

- **Marker Controlled Watershed Segmentation:**

- Calculate distance gradient for edge detection.
- Mark target object using morphological techniques.
- Discard other areas.
- Figure 7 shows segmentation results.

- **Marker Controlled Watershed with Masking:**

- Similar to previous method but with additional masking.
- Marks area containing target object.
- Figure 8 displays segmentation results.

- **Segmentation Analysis:**

- Region growing and marker controlled watershed with masking successfully find target objects.
- Marker controlled watershed fails due to surrounding background.
- Table I compares running times.

- **Binarization:**

- Converts pixel values into black and white.
- Compares black and white pixels with threshold value to determine lung condition (normal or cancer).

Lung Cancer Detection using CT Scan Images

Summary-

- **Model Overview:**

- Replaces Gabor Filter with Median and Gaussian filters in preprocessing.
- Implements watershed segmentation for nodule detection.
- Extracts features like area, perimeter, centroid, diameter, eccentricity, and mean intensity.
- Introduces Support Vector Machine (SVM) for nodule classification.

- **Image Preprocessing:**

- Applies Median filter to remove salt and pepper noise.
- Implements Gaussian filter to smooth and remove speckle noise.

- **Segmentation:**

- Utilizes watershed segmentation to identify cancer nodules.
- Can separate touching objects for accurate segmentation.

- **Feature Extraction:**

- Extracts various features for training the classifier.

- **Classification:**

- Employs SVM for classifying nodules as malignant or benign.
- Defines a function to classify data into two classes based on training inputs.

- **Strengths and Weaknesses:**

- Strengths include improved nodule detection accuracy and classification into malignant or benign.
- Weaknesses include not reaching optimal accuracy levels and inability to classify nodules into different stages.

RESULTS

This section provides a more in-depth explanation of the image processing methods that were utilised on the images contained within the dataset that was used for the research presented here. In addition to this, it displays a figures of lung samples of the various types of lungs that are included in the dataset, such as adenocarcinoma, benign, and carcinoma. In this section, the images of each example are displayed, along with the differences between the images before and after the application of each image processing approach.

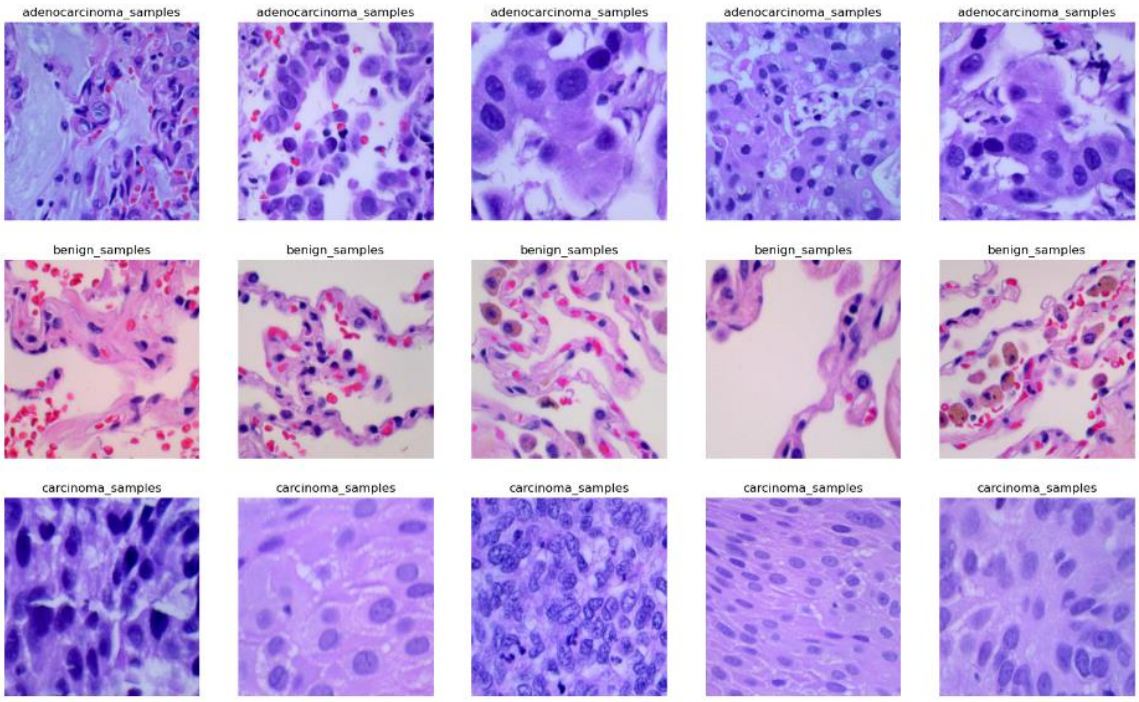

Fig 5: Image of all the samples after reading the images using OpenCV

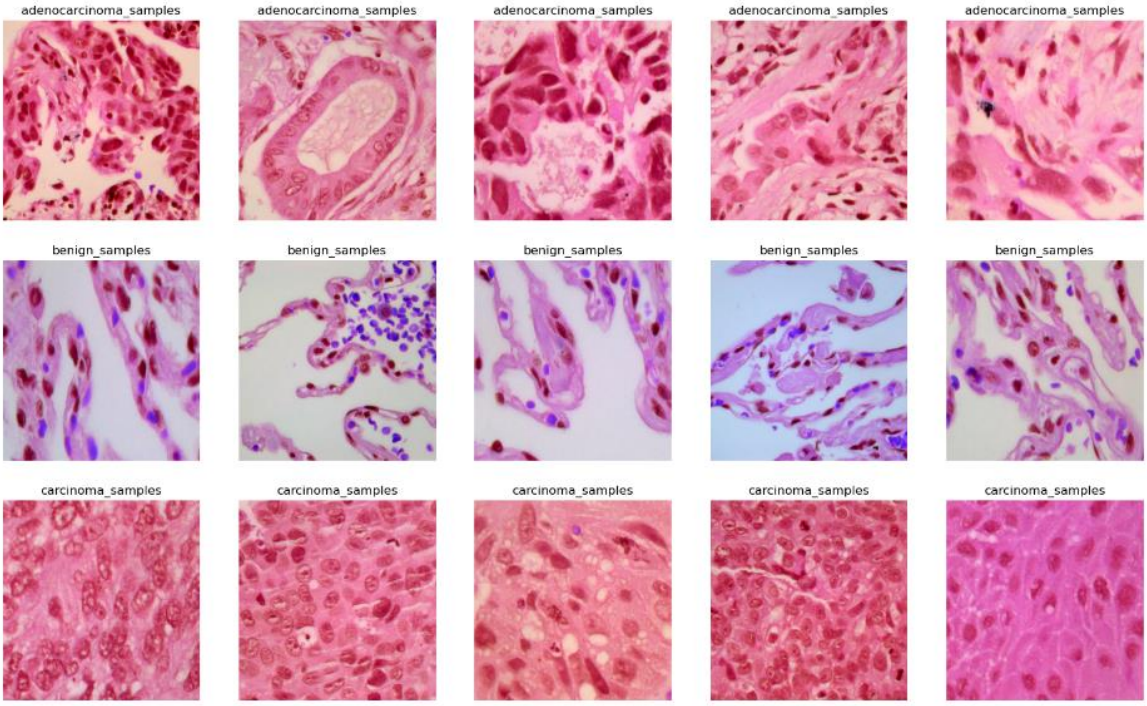

Fig 6: Image of all the samples after color conversion

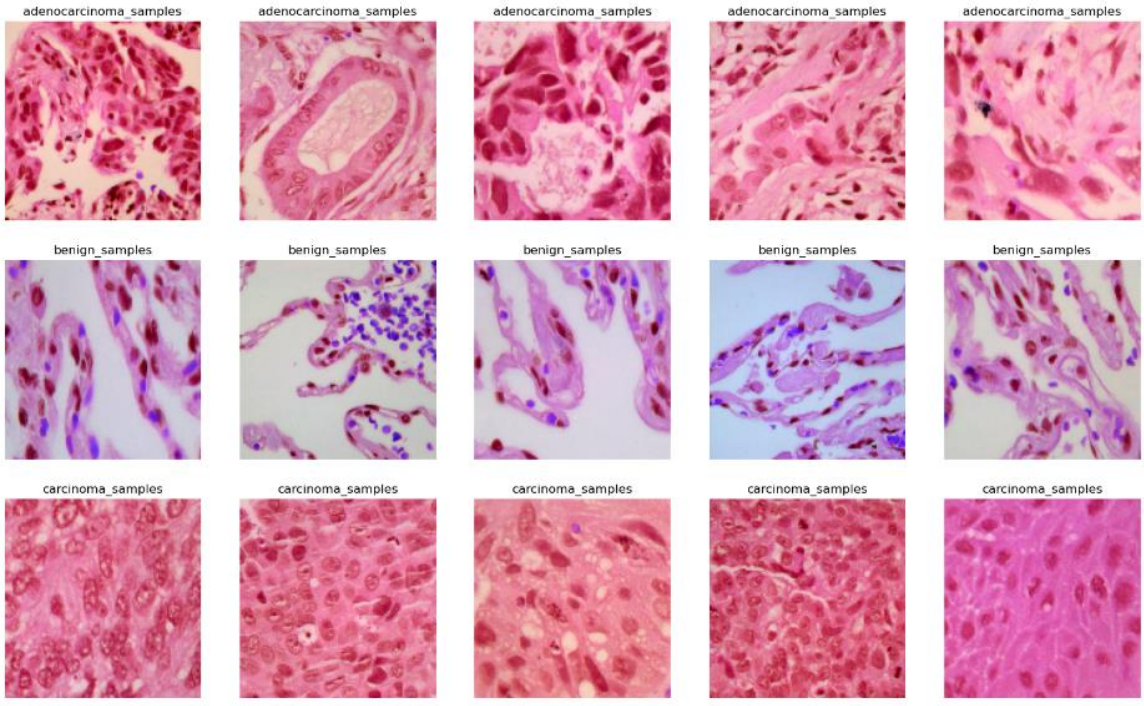

Fig 7: Image of all the samples after resizing image

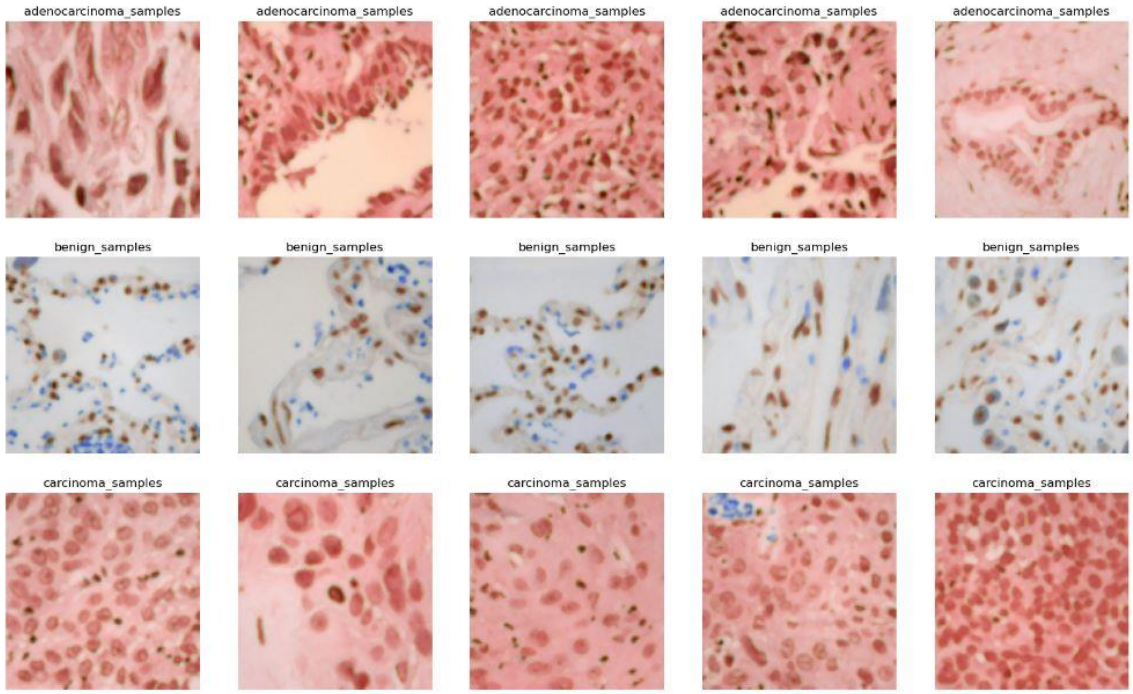

Fig 8: Image of all the samples after applying median filter

From Figure 5, we can deduce that OpenCV initially read the image in the BGR color format. As a result, in order to turn the image into its initial form, a color conversion from BGR to RGB is required, as shown in Figure 6. The necessity to convert the images to a lower dimensional space, such as an image size of 128 by 128 by 3, in order to process the images more quickly can be inferred from Figure 7, which presents the data. After applying this median filter to the images in order to decrease the amount of spatial noise inside the image and further improve the image's quality (as can be seen in Figure 8), the process is complete.

RESULT ANALYSIS

Confusion Matrix

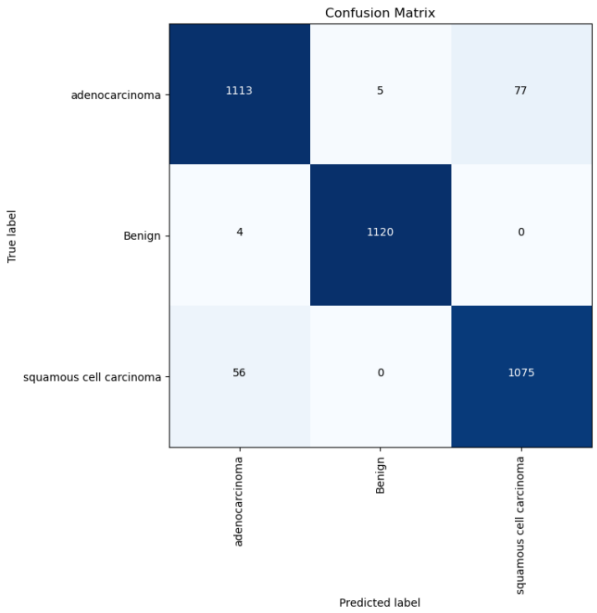

Fig 9: Confusion Matrix of CNN

Fig 10: Confusion Matrix of Efficient B0

CNN Performance curve:

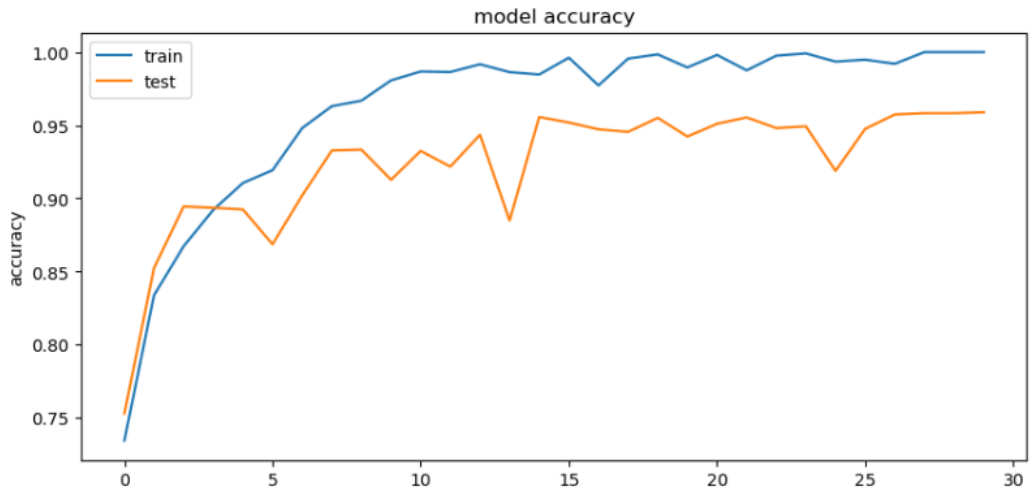

Fig 13: Model accuracy curve of train and test for CNN

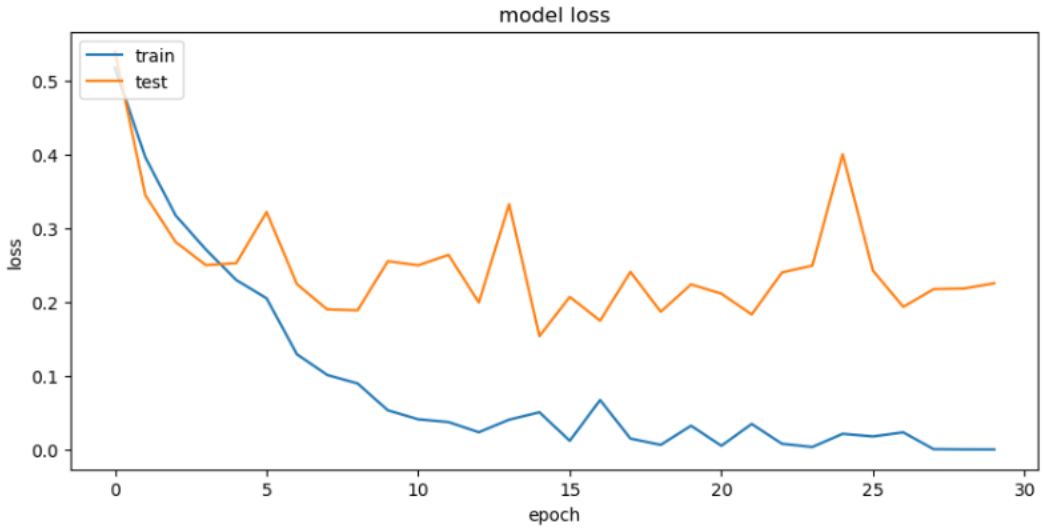

Fig 14: Model loss curve of train and test for CNN

Efficient B0 Performance curve:

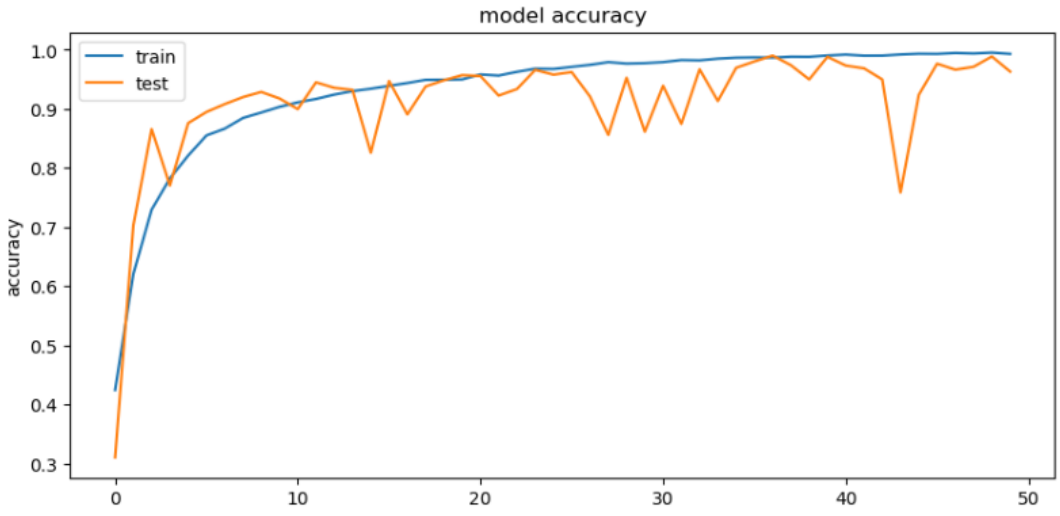

Fig 15: Model accuracy curve of train and test for Efficient B0

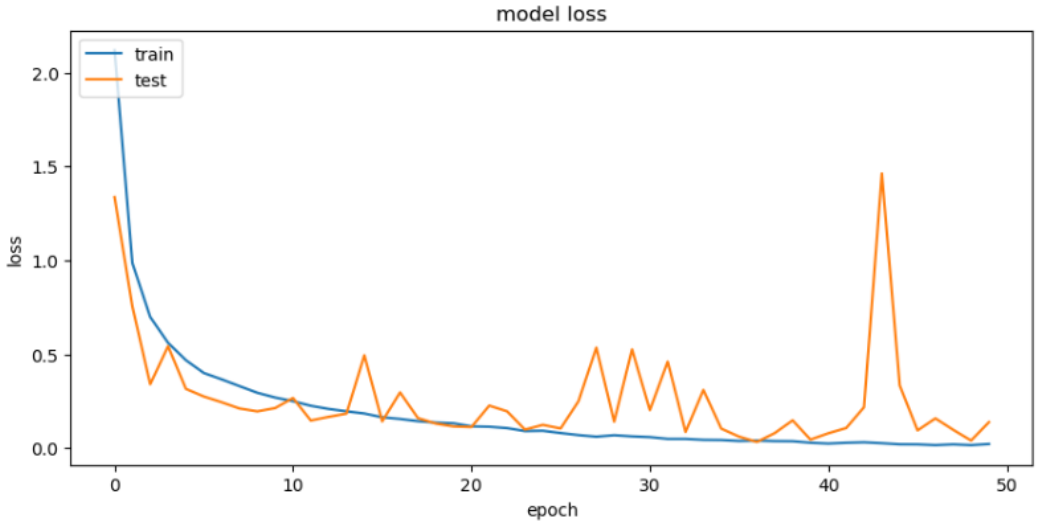

Fig 16: Model loss curve of train and test for Efficient B0

Comparisons:

Models	Accuracy (in %)	Precision	Recall	F1 Score
CNN	95.85	0.959	0.959	0.959
Efficient B0	99.01	0.963	0.963	0.963

Table 1: Algorithms along with performance metrics

Models	Accuracy (in %)		
	Class 1	Class 2	Class 3
CNN	93.14	99.64	95.05
Efficient B0	90.63	99.91	98.76

Table 2: Algorithm performance in terms of accuracy for every grade

References	Models	Accuracy (in %)
Paper-3 (Lung Cancer prediction using machine learning: A comprehensive approach)	ODNN and LDA	94.56

Paper-4 (Multi-stage lung cancer detection and prediction using multi-class svm classifier)	Watershed transform, GLCM and SVM classifier	87
Paper-7 (Automatic lung cancer prediction from chest X-ray images using the deep learning approach)	DenseNet-121	84.02
Paper-8 (Predicting outcomes of nonsmall cell lung cancer using CT image features)	SVM	65
Paper-9 (Lung cancer prediction using feed forward back propagation neural networks with optimal features)	Proposed NN-PSO	97.8
Paper-10 (Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods)	SDS-neural network	89.63
Paper-11 (ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis)	DCNN	93.46
Paper-14 (Lung cancer prediction using robust machine learning and image enhancement methods on extracted	SVM RBF	99.89

gray-level co-occurrence matrix features)		

Table 1: References algorithm along with performance metrics

From Fig 4 and 5 it can be inferred CNN is converging around 20th epoch and after which the model shows a high bias which is observed in both the graphs of accuracy and loss as the training and validation curves starts diverging from each other. Form Fig 6 and 7 it can be inferred that Efficient B0 accuracy increases as the epochs increase but after a certain threshold curve of accuracy and loss for both train and validation tends to diverge which indicates there is a high bias and variance which leads to overfitting. The experimental results of the performance metrics of the different classifiers used in the model are shown in Table I. The results show that the Efficient B0 outperformed the other algorithms with an accuracy of 99.01%, the highest precision rate of 0.963, a recall of 0.963 and an F1 - score of 0.963.

CONCLUSION

This study presented an approach for classifying whether a patient is affected by Lung Cancer along with the severity of the diseases from scans of the patients' blood by using deep learning algorithms. The dataset utilised in the study was obtained from a Kaggle. Various digital image processing techniques like color conversion, image resizing, color conversion and median filtering were used to enhance the images of the dataset before feeding it to a model. The deep learning algorithms utilised were: CNN, Efficient B0. Based on the research results and comparisons between the performance metrics, it was determined that the learning method, Efficient B0 classifier, has outperformed the other algorithms, with an accuracy of 99.01%, a precision rate of 0.963, a recall of 0.963, and an F1 - score of 0.963. The study also indicated that the CNN performed the best for classifying the samples of Class 1 (Adenocarcinoma) with an accuracy of 93 percent and Efficient B0 classified Class 2 (Benign) and Class 3 (Squamous cell carcinoma) best with an accuracy of 99.91 and 98.76 percent. In future, additional research can be done on finding more suitable processing techniques or by creating custom classifiers for achieving a higher accuracy in predicting lung cancer. It can also be focused on classifying individual classes with high accuracy and then combining the best performing models to achieve a better overall accuracy in classifying the target outcome.

REFERENCES

[1] Rahane, W., Dalvi, H., Magar, Y., Kalane, A., & Jondhale, S. (2018, March). "Lung cancer detection using image processing and machine learning healthcare." In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-5). IEEE.

[2] Seibert, R. M., Ramsey, C. R., Hines, J. W., Kupelian, P. A., Langen, K. M., Meeks, S. L., & Scaperoth, D. D. (2007). "A model for predicting lung cancer response to therapy." International Journal of Radiation Oncology* Biology* Physics, 67(2), 601-609.

[3] Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020, March). "Lung Cancer prediction using machine learning: A comprehensive approach." In 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA) (pp. 108-115). IEEE.

[4] Alam, J., Alam, S., & Hossan, A. (2018, February). "Multi-stage lung cancer detection and prediction using multi-class svm classifier." In 2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2) (pp. 1-4). IEEE.

[5] Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., ... & Xiao, G. (2017). "Comprehensive computational pathological image analysis predicts lung cancer prognosis." Journal of Thoracic Oncology, 12(3), 501-509.

[6] Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R., Jawarneh, M., & Asenso, E. (2022). "Lung cancer classification and prediction using machine learning and image processing." BioMed Research International, 2022.

[7] Ausawalaithong, W., Thirach, A., Marukatat, S., & Wilaiprasitporn, T. (2018, November). "Automatic lung cancer prediction from chest X-ray images using the deep learning approach." In 2018 11th Biomedical Engineering International Conference (BMEiCON) (pp. 1-5). IEEE.

[8] Hawkins, S. H., Korecki, J. N., Balagurunathan, Y., Gu, Y., Kumar, V., Basu, S., ... & Gillies, R. J. (2014). "Predicting outcomes of nonsmall cell lung cancer using CT image features." IEEE access, 2, 1418-1426.

[9] Senthil, S., & Ayshwarya, B. (2018). "Lung cancer prediction using feed forward back propagation neural networks with optimal features." International Journal of Applied Engineering Research, 13(1), 318-325.

[10] Shanthi, S., & Rajkumar, N. (2021). "Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods." Neural Processing Letters, 53(4), 2617-2630.

[11] Suresh, Supriya, and Subaji Mohan. "ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis." Neural Computing and Applications 32.20 (2020): 15989-16009.

[12] Wang, Shidan, et al. "Artificial intelligence in lung cancer pathology image analysis." Cancers 11.11 (2019): 1673.

[13] Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.

[14] Hussain, Lal, et al. "Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features." Applied Sciences 12.13 (2022): 6517.

[15] Yu, Kun-Hsing, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nature communications 7.1 (2016): 12474.

[16] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand and S. M. Mastorides, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," arXiv:1912.12142v1 [eess.IV], 2019.

ROLES

1) Jiya Jiveesha (21BCE1255): 8 Literature Reviews, Convolutional Neural Network (CNN), Proposed Architecture, Abstract, Results
2) Meela Akshaya (21BCE1987): 8 Literature Reviews, Image processing, Efficient B0, Introduction, Result Analysis, Conclusion