LUNG CANCER CLASSIFICATION AND PREDICTION USING MACHINE LEARNING AND IMAGE PROCESSING

Rahimunnisa K ,
Department of Electronics and Communication Engineering
Easwari Engineering College, India.
rahimunnisa.k@eec.srmrmp.edu.in

Dhanush S
Department of Electronics and Communication Engineering
Easwari Engineering College, India.
dhanushs0612@gmail.com

Gokul M
Department of Electronics and Communication Engineering
Easwari Engineering College, India.
310620106027@eec.srmrmp.in

Abstract— **Lung cancer has the potential to be life-threatening. Detecting cancer remains a significant hurdle for medical professionals, with the complete understanding of its origins and optimal treatment still elusive. However, timely identification of cancer can greatly enhance treatment prospects. Image processing techniques, including noise reduction, feature extraction, identification of affected areas, and potentially correlating with medical records pertaining to lung cancer history, are employed to pinpoint regions of the lung affected by the disease. This research demonstrates the precise classification and prediction of lung cancer through the utilization of machine learning and image processing technology. Computed Tomography images are utilized for identifying the lung cancer. A computerized tomography (CT) scan utilizes multiple X-ray images captured from various angles around the body, employing computer processing to generate cross-sectional images (slices) that reveal the internal structures, including bones, blood vessels, and soft tissues. A dataset containing thousands of high-resolution lung scans, gathered from Kaggle. The preprocessing phase transforms raw data into a usable format, while a deep learning algorithm assigns significance to the data. In the final stage, a Convolutional Neural Network (CNN) is employed to determine the health status of the lung, distinguishing between normal and abnormal conditions.**

***Keywords*— Lung cancer, Computed Tomography, Machine learning, Deep Learning**

I. INTRODUCTION

Lung cancer is a type of cancer that starts in the lungs. It is one of the most common cancers worldwide and is a leading cause of cancer-related deaths. There are two main types of lung cancer: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is the most common type, accounting for about 85% of lung cancer cases, while SCLC accounts for the remaining 15%.Lung cancer is a disease characterized by the uncontrolled growth and proliferation of abnormal cells within the lung tissue. These cells deviate from normal cellular function due to DNA mutations induced by various factors such as smoking, exposure to air pollutants, and genetic predispositions. According to the American Cancer Society, lung cancer accounts for approximately 14% of all newly diagnosed cancers. In 2018, it was estimated that there were approximately 234,030 new cases of lung cancer in the United States, resulting in approximately 154,050 deaths. Presently, lung cancer surpasses prostate, colon, and breast cancers combined as a leading cause of mortality [2]. Lung cancer is a prevalent and often fatal condition, claiming an estimated 422 lives worldwide each day. Predominantly afflicting individuals over the age of 50, the incidence of lung cancer continues to rise steadily. Due to its challenging detectability compared to other ailments, lung cancer stands as a leading cause of mortality. The primary impediment lies in the minute size of the initial lesion, commonly referred to as a nodule. Initially characterized by diminutive cancer cell dimensions, these lesions progressively evolve into malignancy over time. Hence, early detection plays a pivotal role in disease management. Timely identification significantly enhances survival rates. Recently, advancements in computer vision technology have yielded sophisticated networks capable of autonomously discerning and delineating healthy and tumorous regions [1].

The principal etiological factor of lung cancer is tobacco smoke, encompassing both direct inhalation and exposure to second hand smoke. Additional risk elements include exposure to radon gas, asbestos, environmental pollutants, and certain genetic predispositions.Clinical manifestations of lung cancer encompass persistent cough, hemoptysis, thoracic discomfort, dyspnea, wheezing, laryngeal changes, unexplained weight loss, and generalized fatigue. However, in its incipient stages, lung malignancy may remain asymptomatic, posing challenges to timely detection.

Diagnostic modalities commonly employed include radiographic imaging techniques such as chest Xrays,zcomputed tomography (CT) scans, and positron emission tomography (PET) scans, coupled with histopathological biopsy for definitive identification of malignant cells. Therapeutic interventions for lung cancer are contingent upon the histological subtype and disease stage, comprising surgical resection, chemotherapy, radiotherapy, targeted molecular therapy, immunotherapy, or multimodal approaches.

Early detection assumes paramount importance in enhancing the prognostic outlook for lung cancer, as it facilitates intervention at more amenable disease phases. Screening protocols employing low-dose CT scans are advocated for select high-risk cohorts, notably individuals with a history of smoking, to facilitate early detection and optimize treatment outcomes.

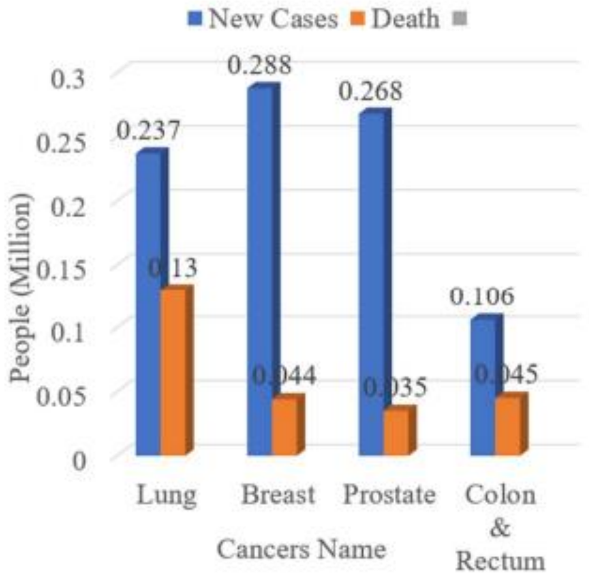


Figure 1: Cancer in 2022 (New cases against death) [5]

Figure 1 illustrates anticipated statistical insights for several cancer types as of 2022, derived from data provided by the American Cancer Society (ACS) [3]. According to ACS, lung cancer exhibits the highest mortality rate among all cancers, estimated at approximately 0.13 million worldwide. Each year, a substantial number of new cases emerge, with projections indicating around 0.237 million cases in 2022. The mortality rate remains notable due to late-stage diagnoses and the relatively high ratio of new cases to mortality, surpassing that of other cancers.

II. WORKING METHODOLOGY

A. Data Collection

Deep learning algorithms can identify affected nodules in lung cancer by analyzing the size, shapes, textures, and intensities of highly detailed images provided by CT scans. 3D CT images, offering a complete imaging of lung capacity, provide a more comprehensive examination of the lungs than their 2D counterparts. Here the input data are consists of three type's especially benign case, malignant case and Normal case. Now see about that image in Detail for Lung cancer detection.

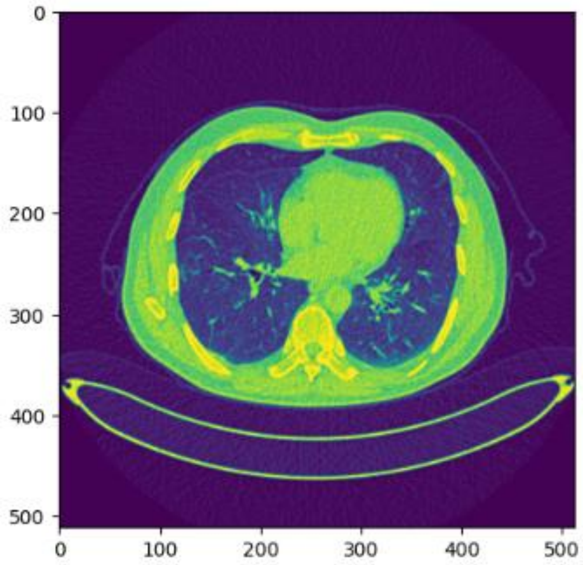


Figure 2: Benign Case Lung Cancer image

Benign lung tumors are growths that do not spread (metastasize) to other parts of the body and are typically not life-threatening. These tumors are often discovered incidentally during imaging tests conducted for other purposes, such as an x-ray. The causes of most types of benign lung tumours are unknown. Some risk factors include: Genetics, infection, smoking. Benign lung tumors typically exhibit small sizes (less than 3 centimetres or roughly 1.5 inches), smooth and regular shapes with borders, and may have variable doubling times, which can be either fast or slow.

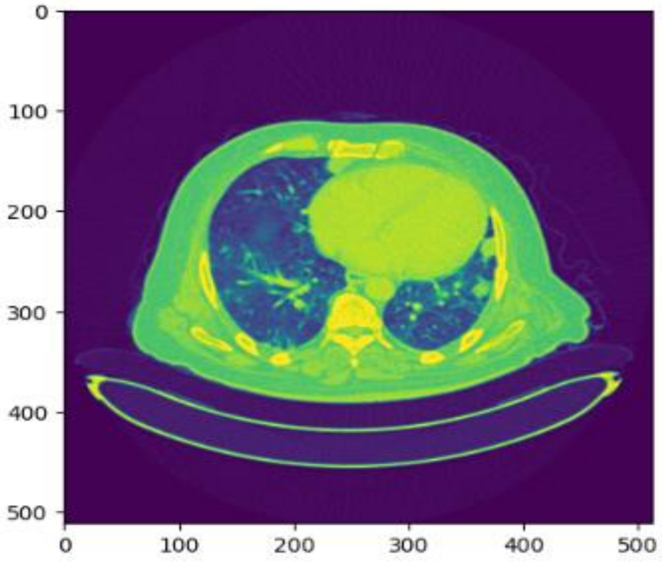


Figure 3: Malignant case lung cancer image

Malignant lung cancer CT images typically show tumors that are small, with sizes less than 3 centimeters (roughly 1.5 inches), and exhibit irregular shapes and borders. The doubling time of these tumors can vary, ranging from fast to slow growth rates.

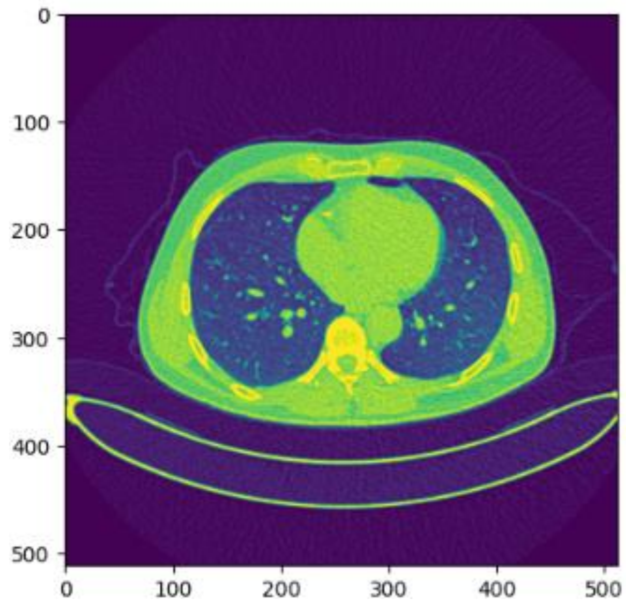


Figure 4: Normal Case Lung cancer image

B. Data Pre-processing

Preprocessing in lung cancer detection images involves several steps to enhance the quality and usability of the data. This typically includes tasks such as noise reduction, normalization of pixel values, resizing or rescaling of images to a standardized resolution, and sometimes augmentation techniques to increase the diversity of the training dataset. Additionally, contrast enhancement may be applied to improve the visibility of subtle features within the images. Overall, preprocessing aims to optimize the input data for more accurate and efficient analysis by machine learning algorithms in detecting lung cancer.

C. Model Architecture

Dense Net, a densely connected convolutional neural network architecture, is effectively employed in lung cancer detection using CT images. Its unique connectivity pattern, where each layer receives direct input from all preceding layers, facilitates feature reuse and enhances gradient flow. This architecture excels in learning intricate patterns and features from CT scans, enabling accurate identification of lung cancer nodules. By leveraging DenseNet's capabilities, researchers have developed robust systems for automated lung cancer detection, aiding in early diagnosis and treatment planning.

D. Training and Validation

During the training process, our model is fed pre-processed CT scan images of the lungs, forming the input for our Convolutional Neural Network (CNN) architecture. Through iterative back propagation, the model adjusts its internal parameters, progressively learning discriminative features and patterns indicative of lung cancer. To mitigate over fitting, we employ a separate validation dataset, distinct from the training data, to monitor the model's performance. This rigorous methodology ensures that our CNN-based model proficiently extends its capacity to identify lung cancer in previously unobserved CT scan images.

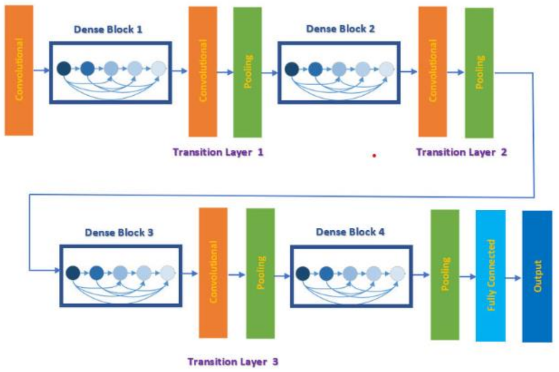


Fig 5: Dense Net architecture

E. Deployment and Integration

The final stage of our methodology focuses on the implementation of the trained Dense Net model for practical clinical applications in the detection of lung cancer.

III.LITERATURE SURVEY

Bhatia et al.[3] A preprocessing pipeline leveraging UNet and ResNet models has been devised by researchers to enhance feature extraction from potentially cancerous lung regions within CT images. Subsequently, an ensemble approach combining XGBoost and random forest classifiers is employed to assess the likelihood of malignancy in CT scans. By aggregating predictions from each classifier, the final determination of malignancy likelihood is made. The methodology yields an 84% increase in accuracy compared to conventional techniques, as demonstrated by evaluation on the LIDC-IRDI dataset.

Joon et al. [4] The study employed an active spline model for lung cancer segmentation analysis, generating X-ray images of the lung. Initial preprocessing recommended the use of a median filter for noise reduction. Subsequently, segmentation was performed utilizing K-means and fuzzy C-means clustering to capture pertinent features. Feature extraction culminated post-segmentation of X-ray images. Classification utilized a Support Vector Machine (SVM) model, implemented within MATLAB to simulate cancer detection outcomes. The study aimed to detect and classify lung cancer using both normal and malignant images.

Faruqui et al. [6] A deep-CNN-based model has been developed to enhance Computer-Aided Diagnosis (CAD) of lung cancer by integrating CT-scan images with wearable sensor-based medical Internet of Things (MIoT) data, thus augmenting diagnostic capabilities. Named LungNet, this model employs a unique 22-layer CNN architecture to extract features from both data sources, achieving a notable accuracy of 96.81% and a low false positive (FP) rate of 3.35% when classifying lung cancer into five classes. LungNet surpasses similar CNN-based classifiers in performance. Furthermore, it accurately classifies stage-1 and stage-2 lung cancers into subclasses with 91.6% accuracy and a FP rate of 7.25%. Trained on a balanced dataset of 525,000 images and operating from a centralized server, LungNet's high accuracy, low FP rate, and ability to perform substage classification position it as a promising solution for automated lung cancer diagnosis systems.

Hasan and Al Kabir [8] The developed algorithms aimed at determining the spread of cancer in patients' lungs, utilizing image processing techniques and statistical learning methods. Evaluation on a dataset of 198 images from the Kaggle platform yielded an accuracy of approximately 72.2%. In contrast, the approach detailed in this article achieves a notably higher accuracy of 99.42%. Additionally, the proposed algorithm demonstrates remarkable performance, with recall, precision, and F-score reaching 99.76%, 99.88%, and 99.82%, respectively. These results underscore the superiority of the method outlined in this study.

Lakshmanaprabu et al. [9]The development of OODN (Optimal Deep Neural Network) involved reducing the number of features in lung CT scans and comparing its performance to other classification algorithms, resulting in a more accurate method. Adoption of automated classification for lung cancer has streamlined human labeling, reducing time and eliminating labeling errors. The study found significant improvements in machine learning algorithms' accuracy and precision in detecting normal and abnormal lung images. Specifically, the research achieved a peer specificity of 94.56%, accuracy of 96.2%, and sensitivity of 94.2% in classifying lung images. These findings demonstrate the feasibility of enhancing cancer detection performance in CT scans, as confirmed by the research results.

IV.OUTCOME

Our proposed lung cancer detection system aims to be efficient, reliable, and suitable for deployment in hospitals. By diagnosing lung cancer at its initial stages, we seek to improve patient care and outcomes.

A. *Advanced Lung Cancer Detection Model*

Our project endeavors to develop an advanced lung cancer detection system based on DenseNet technology. The primary goal is to create a highly accurate and reliable model capable of detecting various types of lung cancer, including benign and malignant cases, with precision.

B. *Improved Data Quality*

Through meticulous data curation and preprocessing techniques such as resizing, normalization, and augmentation, our model ensures the utilization of high-quality input images conducive to effective training. This enhanced data quality enables the model to discern intricate features and patterns specific to lung cancer within medical images.

C. *Efficient Feature Extraction with DenseNet*

The choice of DenseNet architecture is driven by its ability to efficiently extract features through dense connections between layers, allowing for enhanced feature reuse and gradient flow. By leveraging DenseNet's architecture, our model can capture crucial features essential for accurate lung cancer detection.

D. *Real-world Deployment*

The culmination of our efforts involves deploying the trained DenseNet-based model for practical clinical applications. This includes integration into medical imaging systems and telemedicine platforms, facilitating remote consultations and expanding accessibility to lung cancer diagnosis, particularly in underserved regions.

E. *Enhanced Operational Efficiency*

Through automated log collection and analysis, supported by robust log management systems, our project contributes to optimizing healthcare operations. The integration of DenseNet enables early detection of issues and predictive maintenance, reducing downtime and improving resource utilization in medical facilities.

F. *Continuous Monitoring and Updates*

To ensure the sustained performance and relevance of our lung cancer detection model, we establish a system for continuous monitoring and updates. This proactive approach enables the model to remain at the forefront of lung cancer detection technology, adapting to emerging practices and advancements in medical imaging.

G. Model Training and Validation

In our setting a batch size of 8 was used and one epoch takes about 20 seconds. This learning process is quite CPU demanding such that we use up to 50 passes in our training. At the end of 30 epochs, our test accuracy was about 96.8%.

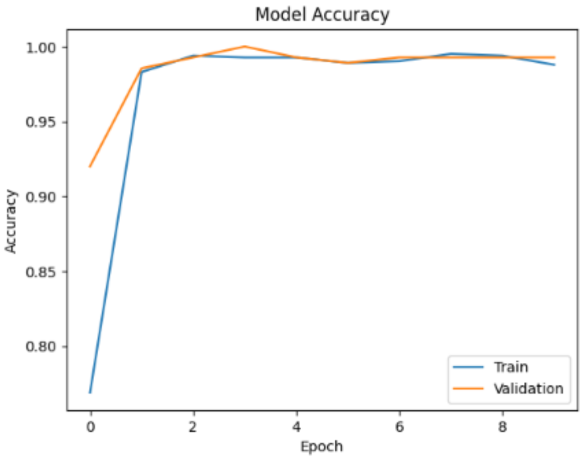


Figure 6: Model accuracy graph

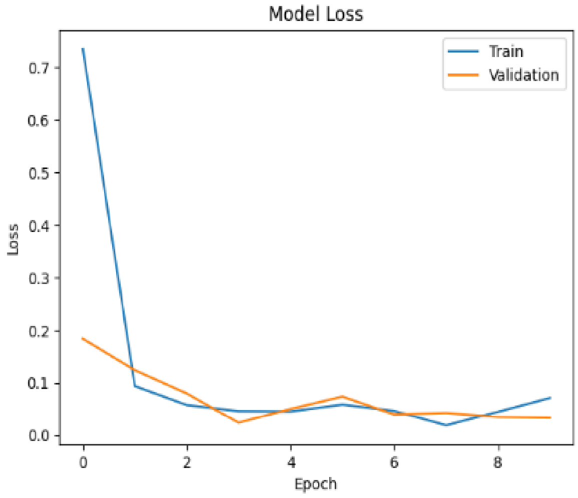


Figure 7:Model loss graph

H. A confusion matrix

A confusion matrix is a matrix that summarizes how well an ML model performs on test data. The most popular application for it is in ranking type models, whose main goal is to predict one target value for each input instance. The various TP, TN, FP, and FN that were derived by using the model on test data are displayed in the above matrix. Below is the confusion matrix that our classification model produced. This matrix, which has three classes denoted as Class 0; Class 1; Class 2; , demonstrates a multi-classification problem. Each row represents the real class, while each column represents the forecast class.

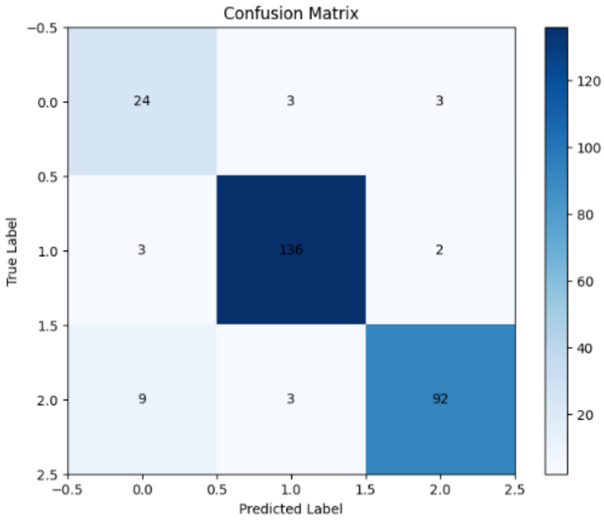


Figure 8: Confusion matrix

V.EVALUATION METRICES

In the context of LC detection, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) each carry distinct meanings. A TP signifies the accurate identification of a particular LC type, TN denotes the correct identification of an image belonging to a different LC type, FP indicates an erroneous identification of an LC type, and FN represents the misidentification of an image belonging to a different LC type.

The accuracy metric evaluates the predictive performance of the proposed model by measuring the proportion of correctly predicted instances, encompassing both TP and TN, relative to the total number of instances in the dataset. Precision represents the proportion of optimistic predictions made by the model. Recall indicates the percentage of TP predictions among all positive cases. The F1-score presents the harmonic mean of precision and recall, where a higher F1-score indicates a model consistently demonstrating elevated levels of recall and precision. Below equation which are used for computing the accuracy, Sensitivity, Specificity for each case.

True Positive (TP): 128

True Negative (TN): 30

False Positive (FP): 0

False Negative (FN): 3

Precision:

Precision refers to the proportion of correctly identified positive samples (True Positives) among all samples classified as positive, regardless of whether they were classified correctly or incorrectly.

$$\text{Precision (P)} = TP / (TP + FP)$$

$$\begin{aligned}\text{Class 0: Precision (P0)} &= TP0 / (TP0 + FP0)\\ &= (30 / (30 + 0)) * 100\\ &\approx 100.00\%\end{aligned}$$

$$\begin{aligned}\text{Class 1: Precision (P1)} &= TP1 / (TP1 + FP1)\\ &= (128 / (128 + 13)) * 100\\ &\approx 90.78\%\end{aligned}$$

$$\begin{aligned}\text{Class 2: Precision (P2)} &= TP2 / (TP2 + FP2)\\ &= (85 / (85 + 19)) * 100\\ &\approx 81.73\end{aligned}$$

$$\begin{aligned}\text{Average Precision} &= P0 + P1 + P2 / 3\\ &=100.00 +90.78+81.73/3\\ &= 90.8\%\end{aligned}$$

Specificity:

Specificity is a metric used in binary classification tasks to measure the ability of a model to correctly identify negative samples. It is calculated as the ratio of true negative predictions (correctly identified negatives) to the total number of actual negative samples. In other words, specificity measures the proportion of actual negative instances that were correctly identified by the model.

$$\text{Specificity} = TN / (TN + FP)$$

$$\begin{aligned}\text{Class 0: Specificity (S0)} &= TN0 / (TN0 + FP0)\\ &= (225 / (225 + 20)) * 100\\ &\approx 91.85\%\end{aligned}$$

$$\begin{aligned}\text{Class 1: Specificity (S1)} &= TN1 / (TN1 + FP1)\\ &= (132 / (132 + 2)) * 100\\ &\approx 98.51\%\end{aligned}$$

$$\begin{aligned}\text{Class 2: Specificity (S2)} &= TN2 / (TN2 + FP2)\\ &= (161 / (161 + 10)) * 100\\ &\approx 94.15\%\end{aligned}$$

$$\begin{aligned}\text{Average Specificity} &= S0 + S1 + S2 / 3\\ &= 91.85+98.51+94.15/ 3\\ &= 95.68\%\end{aligned}$$

Sensitivity:

Sensitivity, also known as recall or true positive rate, is a metric in machine learning that measures the ability of a model to correctly identify positive samples. It is calculated as the ratio of true positive predictions (correctly identified positives) to the total number of actual positive samples. In essence, sensitivity quantifies the proportion of actual positive instances that were correctly identified by the model.

$$\text{Sensitivity (Recall)} = TP / (TP + FN)$$

$$\begin{aligned}\text{Class 0: Sensitivity (R0)} &= TP0 / (TP0 + FN0)\\ &= (30 / (30+ 0))*100\\ &\approx 100\%\end{aligned}$$

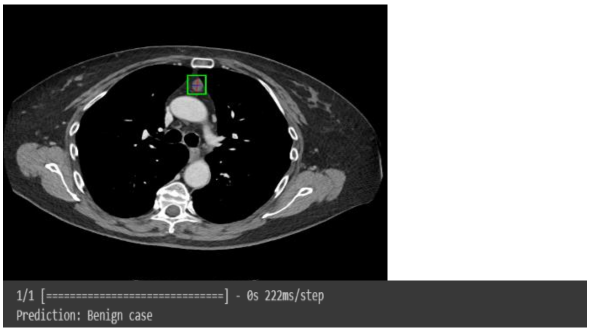
$$\begin{aligned}\text{Class 1: Sensitivity (R1)} &= TP1 / (TP1 + FN1)\\ &= (128 / (128 + 13)) *100\\ &\approx 90.78\%\end{aligned}$$

$$\begin{aligned}\text{Class 2: Sensitivity (R2)} &= TP2 / (TP2 + FN2)\\ &= (85 / (85 + 19))*100\\ &\approx 81.73\%\end{aligned}$$

$$\begin{aligned}\text{Average Sensitivity} &= R0 + R1 + R2 / 3\\ &= 100+90.78+81.73 / 3\\ &= 93.73\end{aligned}$$

VI.RESULTS AND DISCUSSION

In our project, users can input an image file from their system for cancer prediction. The system predicts the type of cancer from three classes: Malignant, Normal, and Benign. The predicted labels are then displayed beneath the provided sample image. This functionality enables seamless prediction and visualization of cancer types based on user-provided images.



```
1/1 [==============================] - 0s 222ms/step
Prediction: Benign case
```

Figure9: Predicted output

VII.CONCLUSION

Our proposed automated classification scheme, based on Dense Net CNN, accurately identifies malignant lung cells in CT images. Evaluation of the method yielded a sensitivity of 96.45%, specificity of 95.52%, and an overall accuracy of 96.3%. Visual analyses further confirm the effectiveness of our approach in identifying malignant cells in CT images.

REFERENCES

[1] Machine Learning-Based Lung Cancer Detection using Multiview Image Registration and Fusion. **Imran Nazir**,[1]Ihsan ul Haq,[1]Salman A. AlQahtani,[2]**Muhammad Mohsin Jadoon**,[3]and Mostafa Dahshan[4.]

[2] Lung Cancer Detection and Classification based on Image Processing and Statistical Learning. Md Rashidul Hasan, Muntasir Al Kabir.

[3] Bhatia S., Sinha Y., Goel L. *Soft Computing for Problem Solving* . Singapore: Springer; 2019. Lung cancer detection: a deep learning approach; pp. 699–705.

[4] Joon P., Bajaj S. B., Jatain A. *Progress in Advanced Computing and Intelligent Engineering* . Singapore: Springer; 2019. Segmentation and detection of lung cancer using image processing and clustering techniques; pp. 13–23.

[5] LCDctCNN:Lung Cancer Diagnosis of CT scan Images Using CNN Based Model. Muntasir Mamun, Md Ishtyaq Mahmud, Mahabuba Meherin, Ahmed Abdelgawad.

[6] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. K. M. Azad, A. Barros, and M. A. Moni, "LungNet: a hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data," *Computers in Biology and Medicine*, vol. 139, Article ID 104961, 2021.

[7] A. Shimazaki, D. Ueda, A. Choppin et al., "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Scientific Reports*, vol. 12, no. 1, Article ID 727, 2022.

[8] M. R. Hasan and M. Al Kabir, "Lung cancer detection and classification based on image processing and statistical learning," *Journal of Emerging Trends in Engineering and Applied Sciences*, vol. 11, pp. 229–236, 2020.

[9] Kulkarni A., Panditrao A. Classification of lung cancer stages on CT scan images using image processing. IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT; 2014; Ramanathapuram, India. 2014. pp. 1384–1388

[10] Gupta A., Awasthi L. K. *GCA* . Las Vegas, Nevada, USA: CSREA Press; 2008. Secure thyself: securing individual peers in collaborative peer-to-peer environments; pp. 140–146.adenocarcinoma. *Pathology* . 2013;45(6):553–558.

[11] Chaudhury S., Shelke N., Sau K., Prasanalakshmi B., Shabaz M. A novel approach to classifying lung cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization. *Computational and Mathematical Methods in Medicine* . 2021;2021:11

[12] Kurkure M., Thakare A. Lung cancer detection using genetic approach. Proceedings -2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA; 2017;