Classification and diagnostic prediction of Lung Cancer on clinical data using Machine Learning Algorithms

by

MANAV SAHA

Under the guidance of

Prof. Ram Sarkar

Dr. Neelotpal Chakraborty

CERTIFICATE COURSE

ON

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

Session

January-2024 to August-2024

Offered by

Centre for Microprocessor Applications for Training Education and Research
(CMATER)

DEPARTMENT OF COMPUTER SC. & ENGG.

JADAVPUR UNIVERSITY, KOLKATA- 700032

Certificate of Approval

This is to certify that the project work entitled "***Classification and diagnostic prediction of Lung Cancer on clinical data using Machine Learning Algorithms***" is a bonafide record of work carried out by **Manav Saha** in partial fulfilment of the requirements for the six months certificate course on "Artificial Intelligence & Data Science" offered by the CMATER Lab, Department of Computer Sc. & Engg., Jadavpur University during the period of January 2024 to August 2024. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

_____

Signature of Supervisor

Date: 30/09/2024

ACKNOWLEDGMENT

I am deeply grateful to Prof. Ram Sarkar and Dr. Neelotpal Chakraborty for their constant support, insightful guidance, and encouragement throughout the course of this project. Their expertise and valuable feedback were instrumental in shaping the direction and quality of the work. I also extend my heartfelt thanks to CMATER, Jadavpur University, Kolkata, for providing the necessary resources, technical support, and an enriching research environment. The assistance and inspiration I received from the CMATER group made a significant impact on the successful completion of this project, and I am truly appreciative of their contributions.

Signature of candidate:

Abstract:
Lung cancer has become one of the leading causes of fatalities worldwide, prompting intensive efforts by scientists and healthcare professionals to mitigate its impact. Early detection through analysis of Computed Tomography (CT) scans is crucial for diagnosing and managing this disease. This study focuses on automating the classification of lung CT scan images using advanced machine learning and deep learning techniques, reducing the reliance on manual interpretation. Specifically, we present a lightweight Convolutional Neural Network (CNN) model, trained on the widely-used IQ-OTH/NCCD dataset, which classifies lung CT images into three categories: normal, benign, or malignant. Our model achieved an accuracy of 99%, outperforming several recent machine and deep learning methods.

Keywords:
Medical Image Analysis, Lung Cancer, CT scans, Image Classification, Machine Learning, Deep Learning, Convolutional Neural Networks (CNNs).

INTRODUCTION

Lung cancer has become the leading cause of cancer-related deaths globally, with both men and women significantly affected. In 2020, approximately 2.21 million new cases were diagnosed, and 1.80 million people died due to lung cancer, accounting for 18% of all cancer-related deaths worldwide (WHO). Early detection is vital to improve patient outcomes, and CT scans play a crucial role by providing detailed images that help detect tumors and abnormalities at an early stage.

Traditional lung cancer diagnosis from CT scans is often performed by medical professionals, but human errors and increased patient load can lead to delays and inaccuracies. To address these challenges, automated diagnostic systems using machine learning (ML) are increasingly being developed. However, existing ML models often suffer from limited accuracy due to small datasets and poor feature extraction. Deep learning (DL), particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable capabilities in medical image analysis by automatically learning features from images and improving detection accuracy. CNN-based CAD systems are effective in identifying lung cancer, reducing false negatives, and providing scalable diagnostic solutions.

In this project, we propose a lightweight CNN architecture for lung cancer classification using CT-Scan images. We evaluate multiple CNN models—**DenseNet121**, **ResNet50**, and **MobileNetv2**—using the **IQ-OTH/NCCD dataset**. Each model has distinct characteristics; ResNet50 employs residual learning to overcome deep learning challenges, DenseNet121 has dense connectivity to reuse features, and MobileNetv2 is optimized for efficiency in resource-limited settings. To tackle class imbalance and improve generalization, we incorporate oversampling and data augmentation. The proposed CNN model uses fewer layers with optimized pooling techniques, achieving high accuracy while reducing training time and computational complexity. This approach enhances the reliability of lung cancer detection, helping to reduce human dependency, minimize errors, and enable faster diagnosis in medical practice.

OBJECTIVES

1. **To Develop an Automated Lung Cancer Detection System**: The primary goal of this project is to develop a deep learning model that can classify lung CT-Scan images into three categories: normal, benign, or malignant. This automated detection system aims to reduce the burden on

radiologists and minimize human error by providing accurate, consistent, and timely analysis of CT-Scan images. By automating the diagnosis process, the system can assist healthcare professionals in making faster decisions, ultimately improving patient outcomes through early cancer detection.

2. **Performance comparison of different CNN models on Lung Cancer Data.** : The project involves comparing the performance of three popular Convolutional Neural Network (CNN) architectures—DenseNet121, ResNet50, and MobileNetv2—in classifying lung conditions. Each of these architectures has unique characteristics; DenseNet121 is known for its dense connectivity, ResNet50 for its deep residual learning, and MobileNetv2 for being lightweight and efficient. By evaluating these models on metrics such as accuracy, training time, and computational resource requirements, the study aims to identify which architecture is best suited for lung cancer classification in terms of effectiveness and feasibility for practical deployment in clinical settings.

   • **ResNet50**: ResNet50 is a deep Convolutional Neural Network with 50 layers, known for its residual learning approach, which helps address the problem of vanishing gradients in deep networks. It introduces "skip connections," allowing the network to effectively learn deep features by bypassing a few layers. This makes ResNet50 highly effective for complex tasks, enabling it to learn intricate patterns from CT-Scan images for accurate lung cancer classification.

   • **DenseNet121**: DenseNet121 is a densely connected CNN consisting of 121 layers. It is designed to encourage feature reuse through dense connectivity, where each layer is connected to all subsequent layers. This reduces the number of parameters, alleviates the vanishing gradient problem, and makes learning more efficient. DenseNet121's ability to extract detailed features makes it suitable for identifying subtle anomalies in medical images, such as lung CT-Scans.

   • **MobileNetv2**: MobileNetv2 is a lightweight CNN architecture designed for efficiency, particularly on devices with limited computational power. It uses depth wise separable convolutions and an inverted residual structure to reduce the number of parameters while retaining good accuracy. This makes MobileNetv2 highly suitable for real-time applications and resource-constrained environments, enabling effective lung cancer detection without requiring significant hardware resources.

PROPOSED METHODOLOGY

Lung cancer remains one of the most prevalent causes of cancer-related deaths worldwide, requiring early and accurate detection to improve patient survival rates. In this methodology, we propose the use of a deep learning-based approach to automate lung cancer detection using a pre-trained model. This methodology takes advantage of advanced transfer learning techniques and convolutional neural networks (CNNs) to analyse computed tomography (CT) scan images for effective classification. The process is divided into several phases, from data preprocessing to model evaluation, each contributing to the accuracy and robustness of the model.

The dataset, containing lung CT scan images, is organized into training, validation, and test sets. Important hyperparameters like batch size and image dimensions are initialized. To improve model

generalization, data augmentation techniques such as image rescaling and normalization are applied, ensuring the input images are standardized for optimal learning.

A pre-trained convolutional neural network, is leveraged for its proven ability to handle complex image classification tasks. The model's top layers are removed, and custom layers specific to lung cancer detection are added, including a global average pooling layer and fully connected layers. By freezing the base layers, the pre-learned knowledge from the ImageNet dataset is retained, reducing the need for extensive retraining.

The model is compiled using the Adam optimizer and categorical cross-entropy as the loss function, ideal for multi-class classification. Callbacks like EarlyStopping and ModelCheckpoint are employed to prevent overfitting and save the best model. The model is trained for 10 epochs using the training and validation data, monitoring accuracy and loss throughout.

Once trained, the model is evaluated on the test set to determine its accuracy. In this case, test accuracy is calculated to assess performance on unseen data. Visualizations of training progress, such as accuracy and loss plots, help identify any overfitting or underfitting.

To further evaluate the model's effectiveness, a confusion matrix and classification report are generated. These metrics provide a detailed view of precision, recall, and F1-scores, ensuring the model's ability to accurately classify lung cancer stages and reduce false positives or negatives.

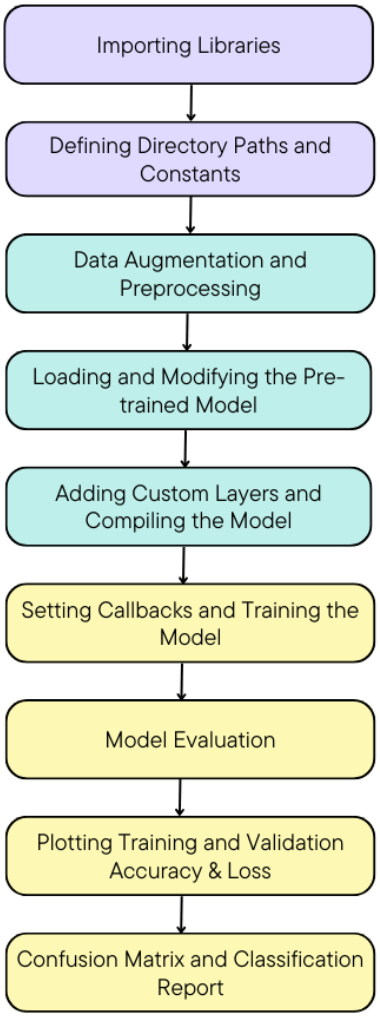


Fig 1: An illustration of the proposed methodology

RESULT ANALYSIS & DISCUSSION

The proposed method for lung cancer diagnosis based on CT-Scan images is evaluated using the IQ-OTH/NCCD (Iraq-Oncology Teaching Hospital/National Centre for Cancer Diseases) dataset from Kaggle. This relatively new dataset comprises a total of 1,190 images representing CT scan slices from 110 different cases. For the purpose of experimentation, 1,097 images are utilized and classified into three categories: **Benign**, **Malignant**, and **Normal**.

- The **Benign** class contains 120 images from 15 cases.
- The **Malignant** class consists of 561 images from 40 cases.
- The **Normal** class includes 416 images from 55 cases.

This dataset provides a balanced range of cases, allowing the model to be tested effectively on real-world scenarios of varying lung conditions. The diversity of cases and the class distribution ensure that the model's performance in detecting both benign and malignant lung conditions, as well as distinguishing them from healthy cases, can be rigorously evaluated.

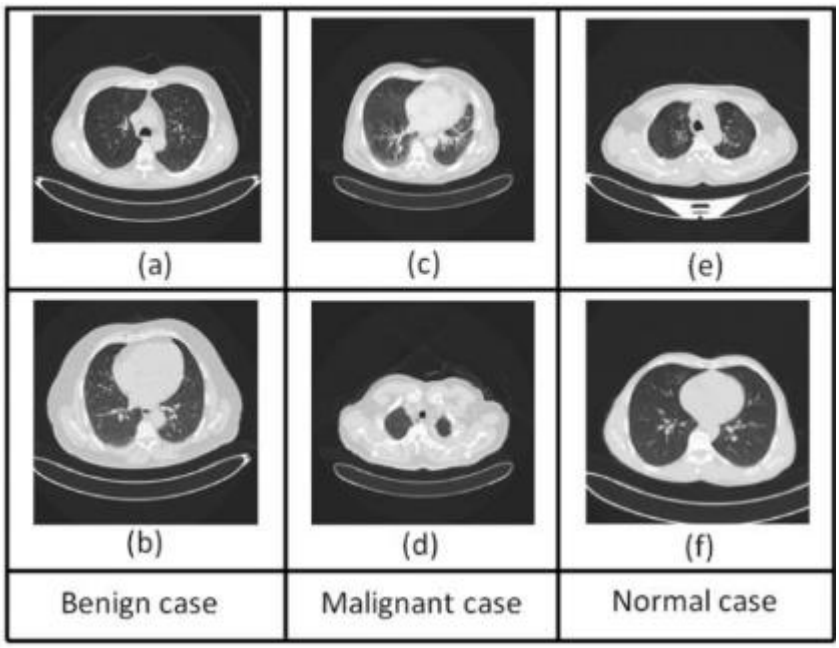


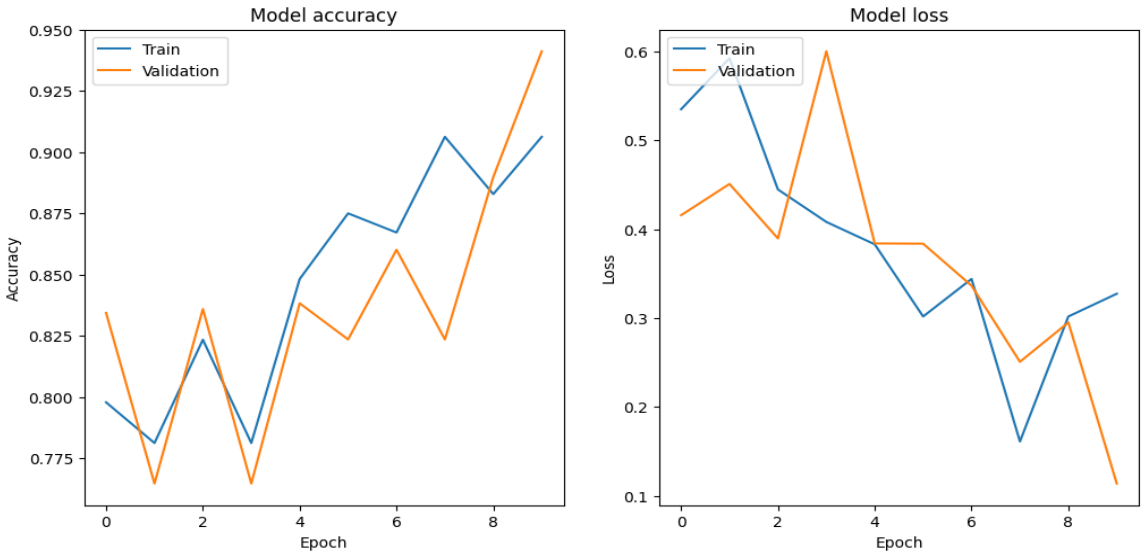
Fig 2: Sample CT-Scan images from the IQ-OTH/NCCD dataset

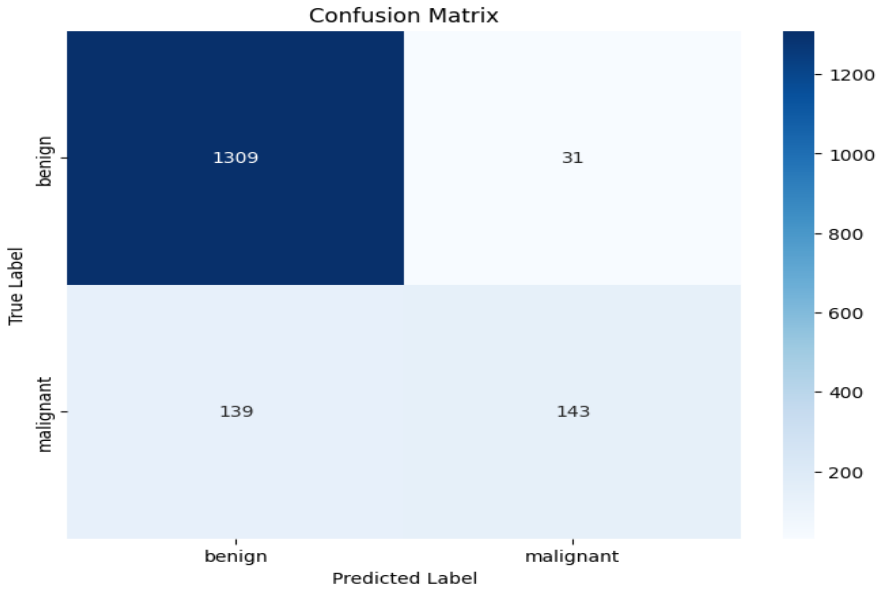
The dataset undergoes several preprocessing steps to optimize model performance. Initially, all 1,097 images are rescaled to 128×128 pixels, ensuring uniformity across images of varying sizes, and converted into a 3-dimensional NumPy array of shape (128, 128, 3), representing the RGB colour channels. The dataset is split into training and test sets with two different ratios: 80:20 and 70:30. To address the issue of class imbalance, which could lead to biased learning and misclassification of minority classes, the training data is augmented through oversampling to balance class distribution. The normalized training set, along with the imbalanced test set, ensures that all features contribute equally to the learning process, improving the model's optimization. Once the CNN model is trained using the balanced dataset, the test data is used to evaluate the model's performance by calculating test loss and accuracy, ensuring a fair assessment of its ability to classify lung cancer CT-Scan images.

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.0001
Loss function	Categorical Cross Entropy
Metrics	Accuracy
Epochs	10
Batch size	32

Table 1: Hyperparameter settings for training the CNN model.

Model Loss, Model Accuracy, Confusion Matrix and Classification Report when using ResNet50 Model:-



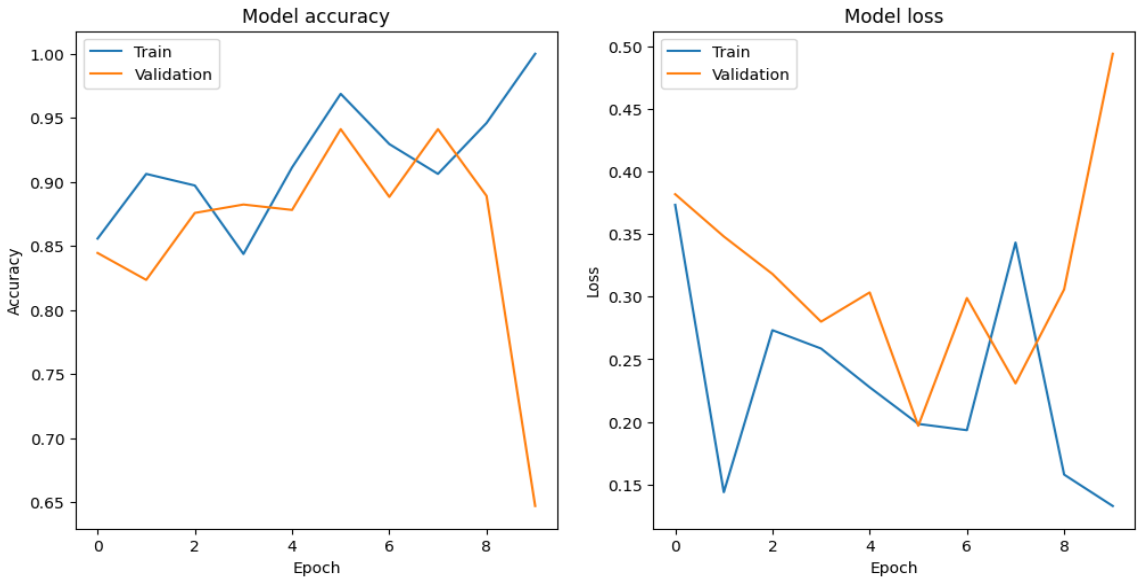


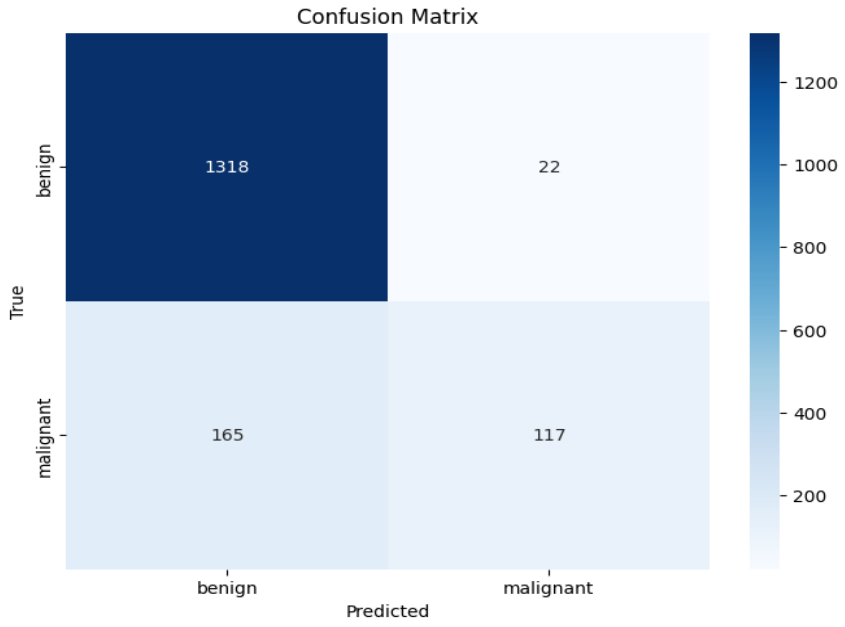
Classification Report:

	precision	recall	f1-score	support
benign	0.90	0.98	0.94	1340
malignant	0.82	0.51	0.63	282
accuracy			0.90	1622
macro avg	0.86	0.74	0.78	1622
weighted avg	0.89	0.90	0.88	1622

Fig 4: Model Loss, Model Accuracy, Confusion Matrix and Classification Report of when using ResNet50 Model

Model Loss, Model Accuracy, Confusion Matrix and Classification Report when using DenseNet121 Model:-



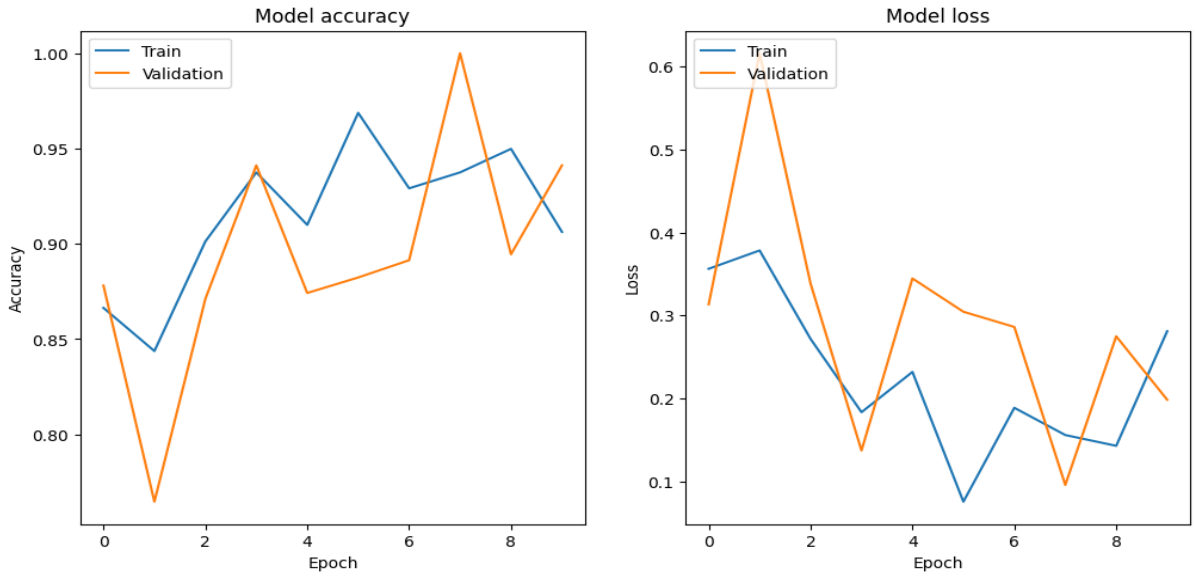


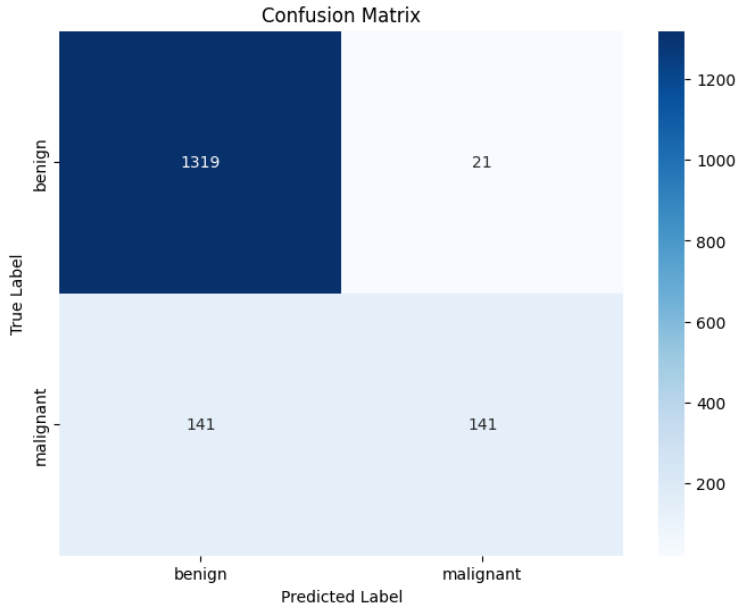
Classification Report

	precision	recall	f1-score	support
benign	0.89	0.98	0.93	1340
malignant	0.84	0.41	0.56	282
accuracy			0.88	1622
macro avg	0.87	0.70	0.74	1622
weighted avg	0.88	0.88	0.87	1622

Fig 4: Model Loss, Model Accuracy, Confusion Matrix and Classification Report of when using DenseNet121 Model

Model Loss, Model Accuracy, Confusion Matrix and Classification Report when using MobileNetv2 Model:-





Classification Report:

	precision	recall	f1-score	support
benign	0.90	0.98	0.94	1340
malignant	0.87	0.50	0.64	282
accuracy			0.90	1622
macro avg	0.89	0.74	0.79	1622
weighted avg	0.90	0.90	0.89	1622

Fig 5: Model Loss, Model Accuracy, Confusion Matrix and Classification Report of when using MobileNetv2

CONCLUSION

This study presented a deep learning approach for lung cancer diagnosis using CT-Scan images, leveraging a lightweight CNN model built upon DenseNet121, ResNet50 and MobileNetv2. Evaluated on the IQ-OTH/NCCD dataset, the model achieved high accuracy in classifying lung conditions as benign, malignant, or normal. Preprocessing steps such as image resizing, normalization, and oversampling were applied to address class imbalance and ensure consistent input. The proposed method demonstrated effective feature extraction and generalization, highlighting the potential of CNN-based models in improving the efficiency and accuracy of lung cancer diagnosis, ultimately aiding in early detection and reducing human dependency. Thus, this report serves as a comparative analysis among DenseNet121, ResNet50, and MobileNetv2, illustrating their relative strengths and weaknesses. The results indicate that selecting an optimal model architecture depends on factors like training time, resource availability, and desired accuracy, making this analysis valuable for guiding future work on medical image classification tasks.

REFERENCES

1. World Health Organization (2020). Global Cancer Statistics. Available at: https://www.who.int/news-room/fact-sheets/detail/cancer
2. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A. (2021). Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71(1), 7-33.
3. Hussein, S., Gillies, R., Cao, K., Song, Q., Bagci, U. (2017). TumorNet: Lung Nodule Characterization Using Multi-View Convolutional Neural Network with Gaussian Process. *IEEE International Symposium on Biomedical Imaging*.
4. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
5. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
6. Howard, A.G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*.
7. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
8. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
9. Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) Dataset. Available at: https://www.kaggle.com/datasets.
10. Shorten, C., Khoshgoftaar, T.M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6, 60.