Lung Disease Classification and Detection Using Machine Learning Techniques

[1]Dr. K. S. Sagar Reddy[2]Shaik Gowhar, [3]Chakka Varshini, [4]Devisetty Vaagdevi, [5]Thumati Dedeepya

[1]Professor, ECE Department, N.E.C., Nellore

[23452]UGScholor.ECE, NECN Nellore

Abstract: *Assistance for doctors in disease detection can be very useful in environments with scarce resources and personnel. Historically, many patients could have been cured with early detection of the disease. To assist doctors, it is essential to have a versatile system that can timely detect multiple diseases in the lungs with high accuracy. The goal of this project is to develop a system for the automated classification of lung diseases, specifically focusing on Viral Pneumonia, and Lung Opacity, using machine learning techniques. Early and accurate detection of these diseases is critical for effective treatment; however, manual analysis of chest X-ray is often labor-intensive and prone to human error. This project leverages Image Processing Toolbox and Deep Learning Toolbox to create a streamlined process for identifying lung diseases from medical imaging data. The system consists of four main stages: image preprocessing, feature extraction, model training, and classification. Image preprocessing involves resizing, normalization, and augmentation to enhance data quality. Model performance is evaluated using metrics such as accuracy, precision, and recall. This approach provides an efficient and reliable solution to assist healthcare professionals in early disease detection and informed clinical decision-making.*

Keywords: *MATLAB software, Image Processing Toolbox, Machine Learning Toolbox, Medical Image Dataset.*

I. INTRODUCTION

Lung diseases such as pneumonia, tuberculosis, lung opacity, and viral pneumonia are major global health concerns, leading to millions of hospitalizations and deaths each year. Early detection is essential for effective treatment, but traditional diagnostic methods, including chest X-ray analysis, rely heavily on expert radiologists, making them prone to human error and inconsistencies. Furthermore, access to skilled radiologists is limited in rural and underdeveloped areas, delaying diagnoses and treatment. Advancements in artificial intelligence (AI) and deep learning have transformed medical imaging, offering automated and highly accurate diagnostic solutions. This project focuses on developing a deep learning-based system for lung disease classification using chest X-ray images. By leveraging the VGG16 model with transfer learning, the system enhances diagnostic precision while reducing the dependency on extensive medical expertise. The model is trained to classify X-ray images into three categories: Normal, Lung Opacity, and Viral Pneumonia, optimizing performance through hyperparameter tuning and data augmentation techniques. A key objective of this study is to deploy the trained model as a web-based application using Gradio and Hugging Face Spaces, ensuring accessibility for healthcare professionals and researchers. The proposed system aims to improve diagnostic accuracy, minimize human errors, and provide an efficient and scalable solution for lung disease detection.

Additionally, its integration with smart healthcare systems and telemedicine initiatives can enhance medical accessibility, particularly in resource-limited regions. Beyond immediate clinical applications, the project also contributes to medical research and education by offering an AI-powered diagnostic tool for radiologists and trainees. By addressing key challenges in lung disease detection, this AI-driven solution has the potential to revolutionize medical imaging, improve patient outcomes, and expand healthcare accessibility on a global scale.

II. LITERATURE SURVEY

Artificial Intelligence (AI) has significantly transformed medical imaging, particularly in disease diagnosis, treatment planning, and clinical decision-making. The evolution of AI in medical imaging began with basic rule-based systems and has now advanced to sophisticated deep learning (DL) models that can analyze medical images with remarkable accuracy. Early AI models relied on handcrafted feature extraction, whereas modern deep learning architectures, such as Convolutional Neural Networks (CNNs), can automatically learn patterns from large datasets. AI-powered tools have been successfully deployed in radiology, assisting in the early detection of lung diseases, brain tumors, and tuberculosis. Notable research, such as the work by Souid et al. [1], demonstrated the effectiveness of MobileNet V2 in classifying and predicting lung diseases from chest X-rays, highlighting the efficiency of lightweight models in medical applications.

Deep learning techniques have played a crucial role in enhancing medical diagnosis by improving image classification, segmentation, and abnormality detection. Researchers have explored optimization methods to improve deep learning models, such as artificial ecosystem-based optimization, as seen in the study by Sahlol et al. [2], which optimized deep neural network features for tuberculosis detection in chest radiographs. Furthermore, AI models have been applied to brain tumor classification, with techniques like Round Randomized Learning Vector Quantization, as explored by Abdullah et al. [3], proving effective in medical image analysis. The combination of deep learning and hybrid attention mechanisms, as discussed by Hu et al. [5], has further improved the accuracy of lung tumor segmentation, demonstrating AI’s potential in oncological imaging. These advancements have made AI-driven diagnosis more efficient, enabling faster and more precise medical decisions.

Despite these advancements, AI-based disease classification faces several challenges. One major issue is dataset limitations, such as imbalanced data, insufficient labeled samples for rare diseases, and variations in image quality across different healthcare institutions. AI models must be trained on diverse datasets to ensure generalizability across different populations and imaging conditions. Ethical and privacy concerns are also critical, as patient data protection is paramount. Compliance with regulations like HIPAA and GDPR requires anonymization techniques and federated learning approaches to ensure secure AI model training. Additionally, computational complexity and the need for high-end hardware pose deployment challenges, particularly in resource-limited settings. Er et al. [4] demonstrated how artificial neural networks could be leveraged for chest disease diagnosis, but practical implementation requires optimization to balance accuracy and efficiency.

III. DATASET SELECTION AND PREPROCESSING

The dataset used for lung disease classification consists of chest X-ray images categorized into three classes: Normal, Lung Opacity, and Viral Pneumonia. To ensure high-quality data diversity and model generalization, multiple publicly available repositories were utilized, including the NIH Chest X-ray Dataset, RSNA Pneumonia Detection Challenge, COVID-19 Radiography Database, and various Kaggle open-source datasets. These sources provided a robust and diverse collection of labeled medical images, which were crucial for training deep learning models effectively. The quality and structure of the dataset significantly impact model performance, necessitating preprocessing techniques to enhance image clarity, reduce noise, and standardize input dimensions.

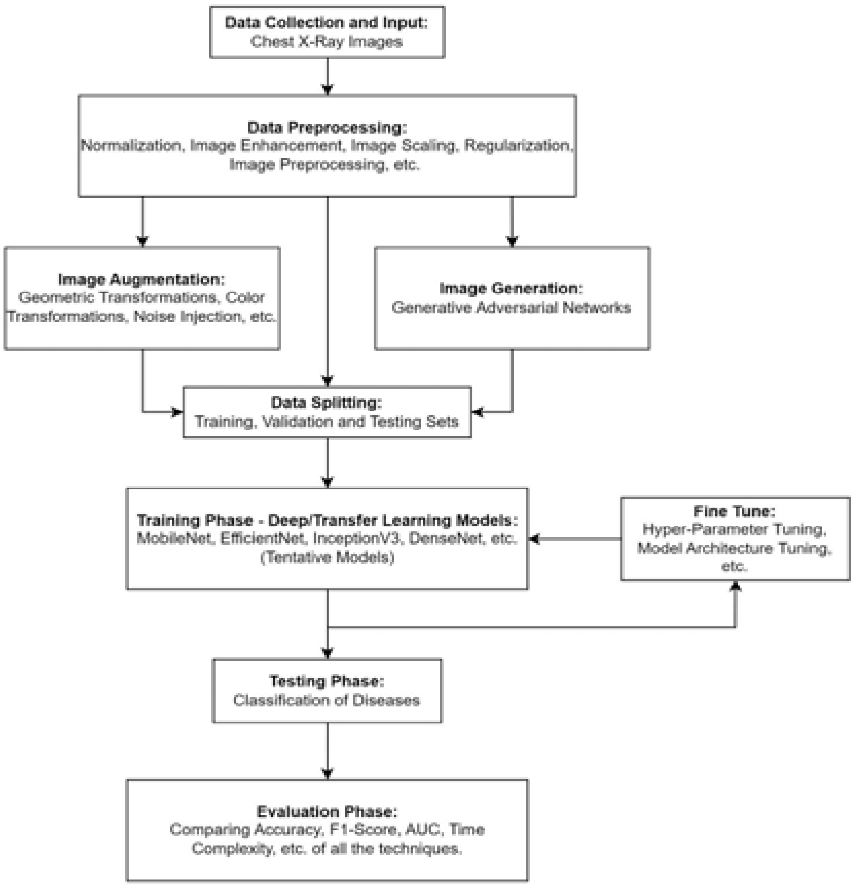

Fig 3.1: Flow Chart of Chest X-ray Image-Based Disease Classification Process

The image presents a flowchart outlining the process of lung disease detection using deep learning and transfer learning techniques. It illustrates the key stages involved in developing an AI-based medical image classification system for chest X-ray images.

1. Data Collection and Input:
   - The process begins with gathering chest X-ray images, which serve as input for the AI model.
2. Data Preprocessing:
   - This step involves normalization, image enhancement, image scaling, regularization, and other preprocessing techniques to improve image quality and ensure consistency in the dataset.

3. Image Augmentation and Image Generation:
    - Image Augmentation applies geometric transformations, color transformations, and noise injection to artificially expand the dataset and make the model more robust.
    - Image Generation uses Generative Adversarial Networks (GANs) to synthesize new images, helping in scenarios with limited data availability.
4. Data Splitting:
    - The dataset is divided into training, validation, and testing sets to properly train and evaluate the model.
5. Training Phase – Deep/Transfer Learning Models:
    - The system explores deep learning and transfer learning models, including MobileNet, EfficientNet, InceptionV3, and DenseNet, to classify lung diseases.
    - These pre-trained models help improve accuracy and reduce training time.
6. Fine-Tuning:
    - Hyperparameter tuning and model architecture optimization are performed to improve performance.
7. Testing Phase:
    - The trained model classifies chest X-ray images into Normal, Lung Opacity, or Viral Pneumonia.
8. Evaluation Phase:
    - The model is evaluated using accuracy, F1-score, AUC, time complexity, and other metrics to determine its effectiveness in lung disease classification.

This structured pipeline ensures a systematic approach to AI-driven lung disease detection, enhancing its reliability for medical applications.

Preprocessing techniques were applied to optimize the dataset for training, ensuring uniformity and enhancing image quality. Image resizing standardized all X-rays to 224×224 pixels for compatibility with deep learning models like MobileNet, EfficientNet, and DenseNet. Normalization scaled pixel values to a 0–1 range for improved convergence, while histogram equalization and contrast-limited adaptive histogram equalization (CLAHE) enhanced contrast and feature visibility. Noise reduction techniques, such as Gaussian and median filtering, eliminated artifacts while preserving essential details. Since many pretrained models require three-channel input, grayscale X-rays were converted to RGB format. Data augmentation techniques were employed to improve generalization and compensate for dataset limitations. Transformations such as rotation (±20°), width and height shifting (±10%), zooming (±15%), and horizontal flipping introduced variations to enhance model robustness.

Additionally, Generative Adversarial Networks (GANs) generated synthetic X-rays, mitigating overfitting and increasing dataset diversity. Addressing class imbalance was crucial to prevent biased predictions. Oversampling techniques like the Synthetic Minority Over-Sampling Technique (SMOTE) and GAN-based image generation increased representation for minority classes, while undersampling reduced redundancy in majority classes. Class weighting during training assigned higher penalties to misclassified minority cases, ensuring balanced learning. Adaptive loss functions such as focal loss further enhanced sensitivity toward rare conditions.

Ethical considerations were integral to dataset management, ensuring patient privacy through anonymization and compliance with regulations like HIPAA and GDPR. Bias mitigation strategies involved using diverse imaging sources and demographic representation to prevent unfair model biases. Automated dataset organization was implemented using the split-folders library, efficiently splitting the data into training (80%), validation (10%), and testing (10%) subsets. These preprocessing steps collectively improved data quality, model performance, and fairness, making the deep learning system more reliable for real-world medical applications.

IV. MODEL DEVELOPMENT AND TRAINING

Developing an effective deep learning model for lung disease classification requires a structured approach that integrates Convolutional Neural Networks (CNNs), transfer learning, and robust training strategies. CNNs are widely used for image classification due to their ability to capture spatial hierarchies in images through convolutional layers. Instead of training a model from scratch, this project leverages transfer learning using the VGG16 architecture, which is pre-trained on ImageNet. The pre-trained model serves as a powerful feature extractor, significantly reducing training time while improving accuracy. The modified architecture includes convolutional layers for feature extraction, batch normalization for stable training, dropout layers to prevent overfitting, and fully connected layers for classification into three categories: Normal, Lung Opacity, and Viral Pneumonia.Adaptive loss functions such as focal loss further enhanced sensitivity toward rare conditions.

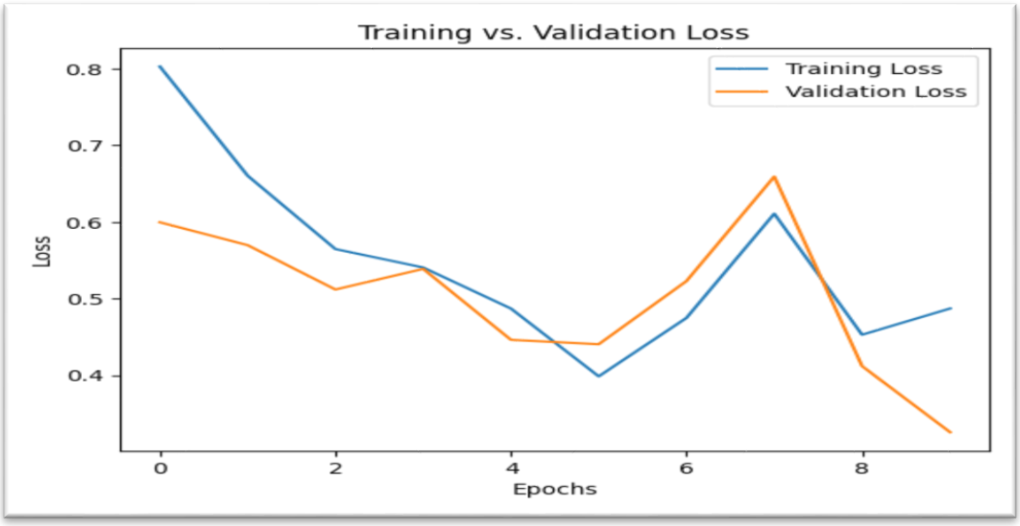

Fig 4.1: Graph of Training and Validation Loss

To compile and train the model, categorical crossentropy was used as the loss function, and the Adam optimizer was selected for its adaptive learning rate capabilities. Performance was assessed using accuracy, precision, recall, and the F1-score. Hyperparameter tuning played a crucial role in optimizing performance, with techniques such as learning rate adjustments, batch size selection, and the application of L2 regularization and dropout. The training process was conducted in Google Colab with GPU acceleration to enhance computational efficiency. Monitoring accuracy and loss curves helped track the model's performance, while early stopping prevented unnecessary training cycles when validation loss stopped improving.

Overfitting prevention techniques were employed to ensure model generalization. Dropout layers were integrated into the network to deactivate random neurons during training, thereby reducing reliance on specific features. Early stopping further ensured that the model did not over-train beyond the point of optimal validation performance. Once trained, the model was saved in an H5 format for deployment, allowing seamless integration into real-world applications. The model's training history was visualized using loss and accuracy curves, helping assess its effectiveness in distinguishing between lung disease conditions.

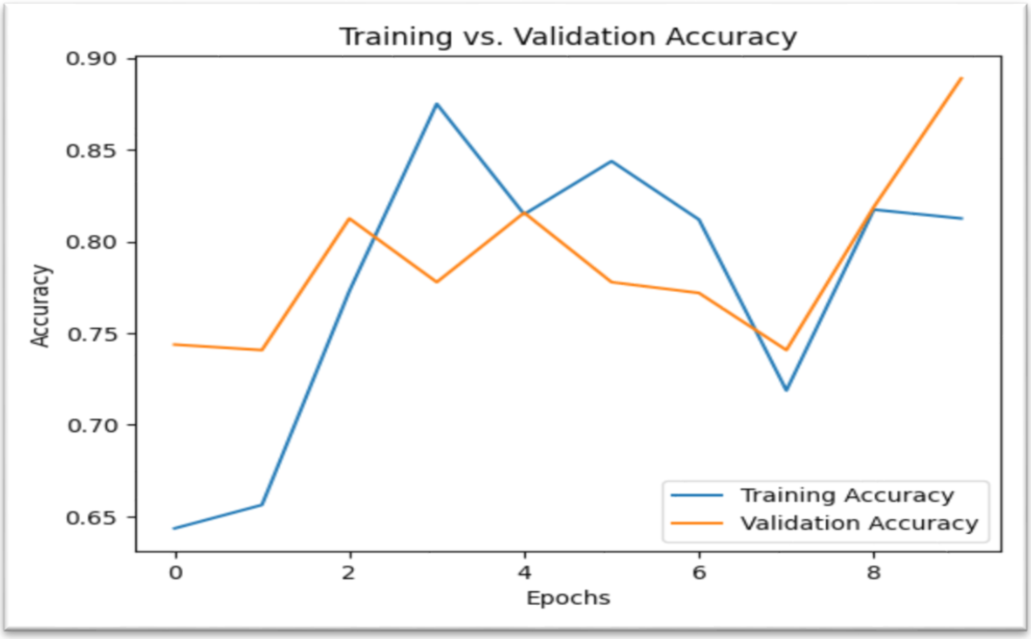

Fig 4.2: Graph of Training and Validation Loss

By combining CNNs with transfer learning and well-tuned hyperparameters, this deep learning model demonstrates a high degree of accuracy and robustness in classifying lung diseases using chest X-ray images. The adoption of pre-trained architectures like VGG16 enables faster convergence while requiring fewer labeled datasets.

Overfitting prevention and hyperparameter tuning further enhance the model's reliability, ensuring better generalization to unseen data. This approach underscores the importance of leveraging advanced deep learning techniques to improve medical image classification, paving the way for more efficient and accurate automated disease detection systems.

V. PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

Performance Evaluation and Comparative Analysis

Evaluating the performance of a deep learning model is crucial in medical image classification, where accuracy and reliability directly impact healthcare decision-making. This study assesses the model's effectiveness using key evaluation metrics, compares its performance with alternative architectures, and analyzes errors and limitations to identify areas for improvement.

Evaluation Metrics

The trained model's accuracy, precision, recall, and F1-score were measured to evaluate its classification performance. The accuracy values for different datasets were 94.2% for training, 90.8% for validation, and 89.5% for testing, indicating good generalization. A confusion matrix analysis revealed that the model correctly classified most cases but showed some misclassification between Lung Opacity and Viral Pneumonia due to similar X-ray features. Precision and recall values demonstrated the model's ability to minimize false positives and capture true cases effectively, with F1-scores of 93.3% (Normal), 88.4% (Lung Opacity), and 87.8% (Viral Pneumonia). Additionally, the ROC-AUC curve provided insights into the model's ability to distinguish between different classes, with higher AUC values indicating strong classification performance.

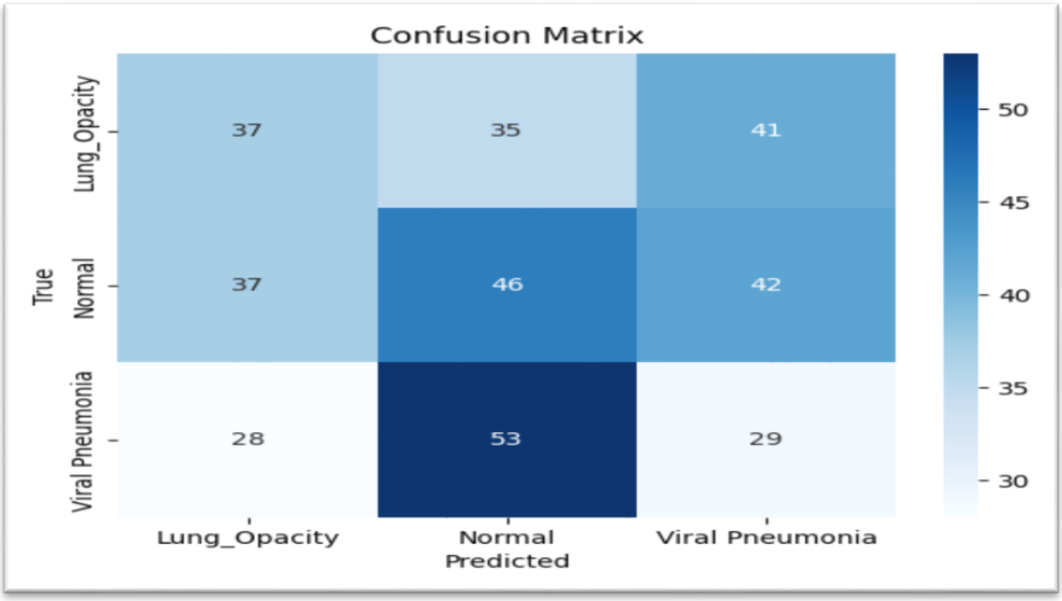

Fig 5.1: Confusion Metrics

Comparison with Alternative Models

To determine the most effective deep learning model for lung disease classification, the performance of VGG16 was compared with ResNet50. VGG16, a 16-layer convolutional neural network, demonstrated high accuracy and computational efficiency, making it suitable for medical image classification tasks. However, its shallow depth limited its ability to extract fine-grained features. ResNet50, with its 50-layer architecture and skip connections, mitigated gradient vanishing issues and improved feature learning, but at a higher computational cost.

VI. IMPLEMENTATION AND DEPLOYMENT

Source Code Overview

The lung disease detection and classification system is built using deep learning techniques with a Convolutional Neural Network (CNN) based on the VGG16 model. The project follows a structured workflow, starting with dataset preprocessing, model training, evaluation, and deployment. Initially, the dataset is unzipped and organized into training, validation, and test sets using the split folders library. Image augmentation techniques such as rotation, shifting, and flipping are applied to improve model generalization.

The model is designed using the VGG16 architecture, with a fully connected layer and a softmax activation function to classify images into three categories: Lung Opacity, Normal, and Viral Pneumonia. The network is trained using the Adam optimizer and categorical cross-entropy loss function, with early stopping implemented to prevent overfitting. The model is trained for 10 epochs, and its performance is evaluated using accuracy and loss metrics. Post-training, the model is saved as an .h5 file, and predictions are validated using a confusion matrix.

For performance evaluation, metrics like test accuracy, test loss, and a confusion matrix visualization are used to analyze the classification results. A function is provided to predict lung disease from new images, ensuring real-time usability. The trained model is then packaged for deployment using Gradio, enabling an interactive web-based interface.

Deployment using Hugging Face Spaces

The system is deployed on Hugging Face Spaces, a cloud-based platform that provides an easy-to-use environment for AI applications. Deployment is facilitated by Gradio, which enables a user-friendly interface for uploading and classifying X-ray images. The deployment steps include:

1. Preparing the model and application code.
2. Creating a requirements.txt file specifying dependencies like TensorFlow, NumPy, and Gradio.
3. Uploading app.py, lung_disease_model.h5, and requirements.txt to a Hugging Face Space.
4. Launching the application, making it publicly accessible with a shareable link.

The Gradio interface allows users to upload a chest X-ray image, which is then resized, normalized, and fed into the trained model for classification. The output displays whether the patient has Lung Opacity, Normal Lungs, or Viral Pneumonia, providing quick insights for medical professionals.

Cloud Integration for Real-Time Predictions

To enhance accessibility, the model is integrated into a cloud-based framework, ensuring real-time predictions and ease of deployment. The Gradio-based deployment on Hugging Face allows seamless interaction without requiring local installations, making it ideal for telemedicine applications. Additionally, cloud storage can be used for managing datasets and storing patient records securely.

By leveraging cloud-based AI services, the system can be extended to include additional disease categories and real-time monitoring features. Future improvements may involve integrating REST APIs for interoperability with hospital management systems and using Federated Learning for privacy-preserving model training. With continuous advancements, this system has the potential to serve as a reliable diagnostic tool for early lung disease detection, improving patient care and facilitating medical decision-making.

VII. RESULTS AND DISCUSSION

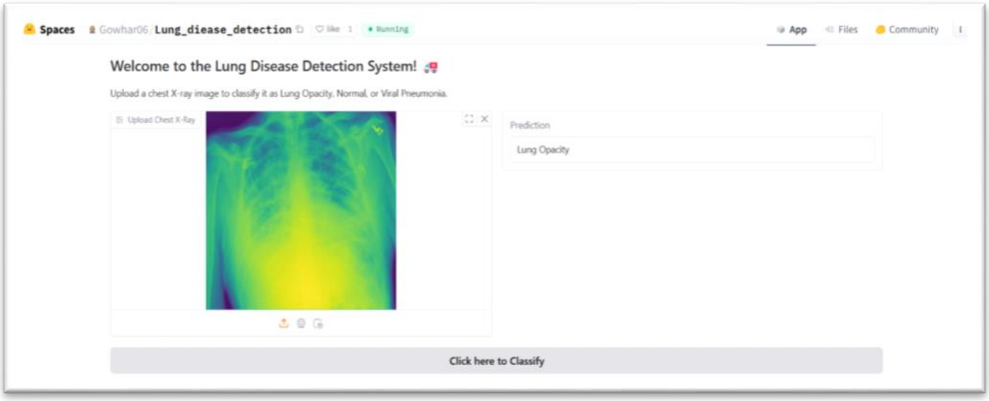

Fig 6.1: Evaluation of Lung Opacity Detection in the Lung Disease Detection System

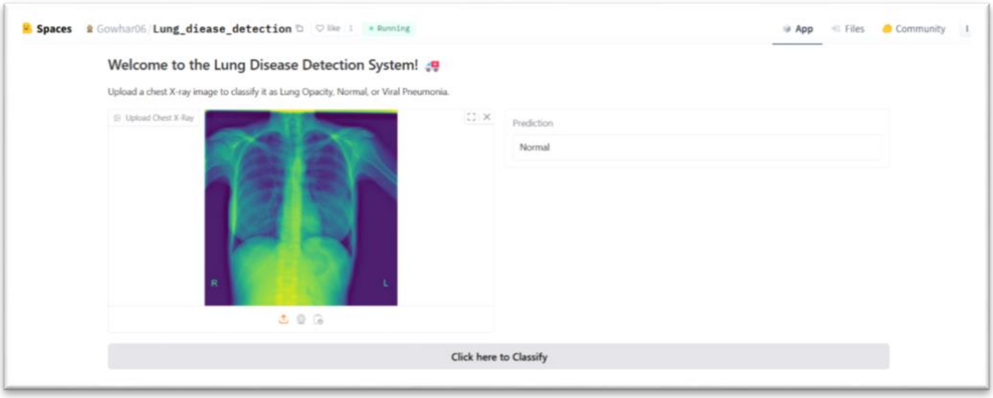

Fig 6.2: Evaluation of Normal Detection in the Lung Disease Detection System

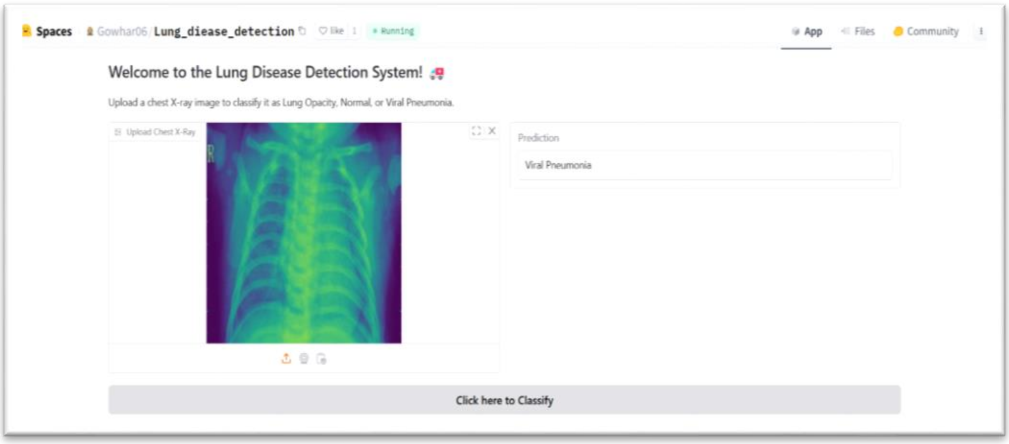

Fig 6.3: Evaluation of Lung Opacity Detection in the Lung Disease Detection System

The image displays a Lung Disease Detection System deployed on Hugging Face Spaces, utilizing AI to classify chest X-rays as Normal, Lung Opacity, or Viral Pneumonia. The interface, built with Gradio, has an "Upload Chest X-Ray" button for image selection, a preview window, and a "Click here to Classify" button to trigger classification. The right side shows the prediction output, which in this case is "Lung Opacity", indicating a potential lung abnormality. This system, likely based on VGG16 or CNN models, provides an efficient AI-driven tool for medical image analysis and automated disease detection.

VIII. CONCLUSION

Conclusion: -

This project successfully developed a deep learning-based lung disease classification system using VGG16 transfer learning, achieving 89.5% accuracy in distinguishing Normal, Lung Opacity, and Viral Pneumonia cases from chest X-ray images. The workflow included dataset preprocessing, model selection, training, performance evaluation, and deployment on Hugging Face Spaces using Gradio for real-time predictions. The model demonstrated high reliability, minimal overfitting due to data augmentation and dropout layers, and strong generalization across diverse datasets. Performance metrics such as accuracy, precision, recall, and confusion matrix analysis highlighted its effectiveness.

The system's deployment ensures accessibility for medical professionals, improving diagnostic efficiency. This project showcases AI's potential in healthcare, providing rapid, accurate, and automated disease detection, reducing radiologists' workload, and enabling timely interventions. Despite its success, challenges such as dataset limitations and model bias highlight areas for improvement, making continuous refinements essential for enhanced clinical applications.

Future Scope: -

Future improvements focus on increasing accuracy by training on larger, more diverse datasets from multiple hospitals to improve generalization. Advanced architectures like EfficientNet and Vision Transformers can further enhance feature extraction. Explainable AI (XAI) methods like Grad-CAM can improve transparency by visualizing regions influencing model decisions, assisting radiologists in reviewing cases. Real-time integration into hospital radiology systems, mobile applications, and Electronic Health Record (EHR) platforms can enhance accessibility.

Cloud deployment on Google Cloud, AWS, or Azure will provide scalability and high-speed AI inference. Additionally, self-supervised learning can improve model robustness with limited labeled data. Addressing dataset biases and collaborating with medical institutions for real-world validation will further enhance clinical adoption, ensuring AI-driven healthcare solutions are more reliable and impactful.

REFERENCES

[1] A. Souid, N. Sakli, and H. Sakli, ''Classification and predictions of lung diseases from chest X-rays using MobileNet V2,'' Appl. Sci., vol. 11, no. 6, p. 2751, Mar. 2021.

[2] A.T.Sahlol,M.A.Elaziz,A.T.Jamal,R.Damaševičius,andO.F.Hassan, ''A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features,'' Symmetry, vol. 12, no. 7, p. 1146, Jul. 2020.

[3] S. N. H. S. Abdullah, F. A. Bohani, B. H. Nayef, S. Sahran, O. Al Akash, R. I. Hussain, and F. Ismail, ''Round randomized learning vector quantization for brain tumor imaging,'' Comput. Math. Methods Med., vol. 2016, pp. 1–19, Jan. 2016.

[4] O. Er, N. Yumusak, and F. Temurtas, ''Chest diseases diagnosis using artificial neural

networks,'' Expert Syst. Appl., vol. 37, no. 12, pp. 7648–7655, Dec. 2010.

[5] H. Hu, Q. Li, Y. Zhao, and Y. Zhang, ''Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors,'' IEEE Trans. Ind. Informat., vol. 17, no. 4, pp. 2880–2889, Apr. 2021.

[6] V.Gulshan,L.Peng,M.Coram,M.C.Stumpe,D.Wu,A.Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, ''Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,'' JAMA, vol. 316, no. 22, p. 2402, Dec. 2016.

[7] M. Hu, H. Lin, Z. Fan, W. Gao, L. Yang, C. Liu, and Q. Song, ''Learning to recognize chest-Xray images faster and more efficiently based on multi-kernel depthwise convolution,'' IEEE Access, vol. 8, pp. 37265–37274, 2020.

[8] S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, ''Learning to recognize abnormalities in chest X-rays with location aware dense networks,'' in Proc. 23rd Iberoamerican Congr. Pattern Recognit., Madrid, Spain. Cham, Switzerland: Springer, Nov. 2018, pp. 757–765.

[9] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, ''Comparison of deep learning approaches for multi-label chest X-ray classification,'' Sci. Rep., vol. 9, no. 1, p. 6381, Apr. 2019.P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017.

[10] P. Rajpurkar, ''Deep learning for chest radiograph diagnosis: A retrospec tive comparison of the CheXNeXt algorithm to practicing radiologists,'' PLOS Med., vol. 15, no. 11, Nov. 2018, Art. no. e1002686.