IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 6, June 2025

Impact Factor: 7.67

Lung Disease Prediction Using Machine Learning Algorithms And GAN

Vishal Borate [1], Dr. Alpana Adsul [2], Palak Purohit[3], Rucha Sambare[4], Samiksha Yadav[5], Arya Zunjarrao[6]

Assistant Professor ,Department of Computer Engineering [1]
Associate Professor, Department of Computer Engineering [2]
Students, Department of Computer Engineering[3-6]
Dr. D. Y. Patil College of Engineering & Innovation Talegaon, Pune, India

Abstract: *Lung diseases, particularly lung cancer, pose a significant global health challenge, ranking among the leading causes of death worldwide. Early and accurate detection is critical for improving patient survival and treatment outcomes. Deep learning techniques, especially convolutional neural networks (CNNs), have shown remarkable success in automating lung disease detection from medical images. However, their performance typically depends on access to large, well-annotated datasets, which are often limited in clinical settings. In this study, we propose a deep learning-based classification framework, Lung-GAN, which combines generative adversarial networks (GANs) with CNNs to enhance detection accuracy of lung cancer using chest CT and X-ray images. Our approach utilizes GANs to generate high-quality synthetic images that augment the training dataset, improving model generalization. We focus specifically on classifying four categories: large cell carcinoma, adenocarcinoma, squamous cell carcinoma, and normal lung tissue. Experimental results show that our framework achieves superior classification performance compared to existing models, even with limited labeled data. This method has the potential to support early diagnosis, reduce radiologist workload, and improve clinical decision-making in lung cancer care*

Keywords: Lung cancer, CT scan, X-ray, Adenocarcinoma, Large cell carcinoma, Squamous cell carcinoma, Deep learning, Generative adversarial networks, Labeled data augmentation.

I. INTRODUCTION

Lung cancer remains one of the most deadly and commonly diagnosed cancers globally, accounting for a significant number of cancer-related deaths each year. Among its subtypes, adenocarcinoma, squamous cell carcinoma, and large cell carcinoma are the most prevalent forms of non-small cell lung cancer (NSCLC), collectively responsible for approximately 85% of lung cancer cases. Early and accurate identification of these subtypes is essential for determining treatment strategies and improving patient outcomes. However, traditional diagnostic methods using chest X-rays and computed tomography (CT) scans can be challenging due to visual similarities between cancer types and the subtle nature of early-stage tumors.

Although chest X-rays and CT scans are indispensable tools in clinical workflows, their interpretation often depends on the subjective assessment of radiologists. This introduces diagnostic variability and delays, especially in regions facing a shortage of skilled professionals. In recent years, deep learning techniques have emerged as a powerful alternative for automating medical image analysis and reducing human dependency. Convolutional Neural Networks (CNNs), in particular, have shown promise in detecting complex visual patterns, offering rapid and consistent results.

Despite these advancements, most deep learning models require large volumes of annotated data to perform well. Creating labeled datasets in the medical domain is resource-intensive, often requiring expert radiologists and significant time investment. Moreover, the availability of balanced, labeled datasets for all lung cancer subtypes remains limited.

To address these challenges, we propose a GAN-based supervised deep learning framework, referred to as Lung-GAN, designed to improve the classification of lung cancer types from CT and X-ray images. Our method leverages a

DOI: 10.48175/IJARSCT-27926

generative adversarial network to augment the labeled training data by generating realistic synthetic images, which are then used to train a CNN classifier. This approach helps mitigate the limitations of small datasets, enhances model generalization, and leads to better diagnostic performance.

Key Contributions:

a) We develop a supervised deep learning model that integrates GAN-based image generation with CNN-based classification for accurate detection of lung cancer subtypes.

b) Our framework specifically classifies four categories: adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and normal lung tissue.

c) We perform extensive evaluation using multiple labeled chest X-ray and CT datasets and demonstrate significant improvements in classification performance over traditional CNN models.

d) We visualize and analyze the learned features to provide interpretability and clinical relevance of the classification results.

II. LITERATURE SURVEY

In paper [1] author focuses on predicting lung diseases using machine learning algorithms, addressing conditions like asthma, COPD, bronchitis, emphysema, and lung cancer. The dataset used in the study includes 323 samples, each with 19 features capturing various symptoms and clinical indicators linked to different lung conditions. To build predictive models, five different machine learning approaches were explored, incorporating both ensemble and probabilistic techniques. Model evaluation was carried out using K-Fold Cross Validation to ensure robust performance assessment. The algorithms achieved precision rates of 88.00% (Bagging), 88.92% (Logistic Regression), 90.15% (Random Forest), 89.23% (Logistic Model Tree), and 83.69% (Bayesian Networks), with Random Forest being the model that execute the finest. This research demonstrates machine learning's potential in early detection and prediction of lung diseases, which could aid in proactive healthcare measures.

In paper [2] author presents a lightweight convolutional neural network (CNN) for diagnosing lung diseases, including COVID-19 and pneumonia, using chest X-ray images. The model achieved 89.89% accuracy on a public dataset, outperforming the Efficient-Net B2 model at 85.7%. While both models effectively identified COVID-19, they struggled to distinguish between viral and bacterial pneumonia due to limited viral pneumonia samples, due to its low computational demands, the model is well-suited for deployment in resource-constrained environments such as IoT-based systems and medical facilities with limited infrastructure. [3] Future enhancements could aim to improve image resolution and mitigate class imbalance issues, thereby boosting the accuracy of viral pneumonia detection.

In paper [4] author proposed a self-operating diagnostic system designed to detect various lung diseases using chest X-ray and CT images. This system integrates a customized Convolutional Neural Network (CNN) with two pre-trained deep learning architectures—AlexNet and VGG16Net. [5] The diagnostic process involves two main stages: first, image pre-processing using a novel enhancement method based on a k-symbol Lerch transcendent function; second, disease classification through deep learning models. Tested on public datasets, the method achieved high accuracy (98.60% for X-Ray, 98.80% for CT), demonstrating the effectiveness of the image enhancement and classification models.

In paper [6] author develops a robust deep learning-based system for identifying lung diseases, including COVID-19, pneumonia, lung opacity, and normal states, from chest X-ray images. Using a Kaggle dataset, the images are preprocessed for contrast enhancement and noise removal, with near-miss resampling addressing dataset imbalance. Ensemble learning techniques with VGG16, InceptionV3, and MobileNetV2 are used, achieving 94% accuracy with a three-level ensemble method. [7] The system includes a web interface for remote access. This research demonstrates the capability of ensemble deep learning for automated, profitable lung disease detection, providing valuable decision support for healthcare, especially in resource-limited settings.

In paper [8] author discusses a disease forecast system that employ machine learning to predict illnesses based on user-entered symptoms. The system not only predicts diseases but also offers health advice, leveraging machine learning models like Random Forest, Naïve Bayes, Decision Tree, and K-Nearest Neighbor (KNN). [9] The prediction process is facilitated through Python and Tkinter for the user interface, using a dataset from hospitals for training. The system evaluates results based on metrics like accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, achieving an overall accuracy of around 95%. It provides a user-friendly interface and stores data for future improvements.[10]

In paper [11] author explores the utilization of machine learning approach for the differential determine of Tuberculosis and Pneumonia, diseases with similar symptoms. [12] The study develops a classification model using Naïve Bayes, Decision Tree, and Random Forest algorithms, with the highest accuracy improvement seen with Naïve Bayes and overall best performance from Random Forest. The results show a slight performance increase with discretization techniques.

[13] Future work aims to extend the model to other diseases with similar symptoms and improve accuracy through additional preprocessing and discretization methods.

In paper [14] author discusses using machine learning and deep learning techniques such as CNN, image processing to develop algorithms for early detection and diagnosis of lung diseases, including asthma, COPD, tuberculosis, pneumothorax, and lung cancer. By analyzing patient data and chest X-ray images, the project aims to create a binary classification model that assists doctors in making timely diagnostic decisions to improve patient outcomes. [15] In future work, this plans to train the model with larger datasets and adjust parameters to enhance processing speed. Additional performance metrics will be evaluated, and pre-trained models may be experimented with to further improve accuracy.

In paper [16] author discusses the employment of CNNs and deep learning for nodule classification and lung cancer prediction from CT imaging, achieving high AUC performance. It emphasizes the need to define how CADx outputs should be integrated into clinical decision-making, addressing questions about risk assessment, algorithm incorporation into guidelines, and comparisons with radiologists. It also emphasizes the importance of considering the characteristics of training and validation datasets, such as smoking history or malignancy.

In paper [17] author develops COPD and Asthma Physiology Score (CAPS) was developed using logistic regression analysis on 8,527 individuals with obstructive airway disease (mean age 65.9 years, hospital mortality 35.5%). With an AUC of 0.718, CAPS, which measures eight different variables (heart rate, MAP, pH, salt, urea, creatinine, albumin, and WBC count), demonstrated fair discrimination. This performance exceeded SAPS II, APACHE II, and APACHE III ratings in 7,957 patients that were validated.

In paper [18] author reviews the limitations of existing lung cancer prediction models for patients being evaluated for surgery and the need for models that address this specific population. By developing the TREAT model, a clinical prediction model for lung cancer. Authors used logistic regression to develop TREAT (Thoracic Research Evaluation and Treatment) model reaching upto 87% accuracy which is measured through AUC (Area under the ROC curve). Also comparing the developed model to existing mayo clinic model which gives 80% of accuracy.

In paper [19] author introduces deep learning system using multi-stream, multi-scale convolutional neural networks to automatically classify pulmonary nodules in CT scans and to process multiple 2D views of nodules across different scales, it also compares against classical machine learning methods like Support Vector Machines (SVMs) using intensity and unsupervised features. Gaining overall accuracy of 79.5% and F-measure (harmonic mean of precision and recall) per class ranged from 43.4% for spiculated nodules to 85.7% for calcified nodules. Although facing an issue of data imbalance, talking about other such existing models like PanCan Model, Lung-RADS Guidelines, KNN (k-nearest neighbors), SIFT (Scale-Invariant Feature Transform).

In paper [20] author converses about segmentation framework for pulmonary nodules in lung CT images, focusing on different types of nodules such as solid, part-solid, and non-solid aiming to improve the segmentation accuracy for each type, considering internal texture and external attachment to pleura or vessels. iterative morphological filtering, vasculature pruning technique, 3D region growing with fuzzy connectivity, Ellipsoid Approximation, Selective [21] Enhancement Filtering addressing the challenges of segmenting different types of pulmonary nodules and introduces a robust framework that adapts based on the nodule type. She also discusses and borrows some ideas from existing models such as Kuhnigk, Moltz, Kubota, Dehmeshki. [22] The proposed method achieved the highest accuracy for both

DOI: 10.48175/IJARSCT-27926

solid and non-solid nodules compared to existing techniques, accuracy of 99% for solid and part-solid nodules and 98% for non-solid nodule.

In paper [23] author outlines the methods used for data collection and validation in the (ICNARC) Case Mix Programme (CMP), which assesses outcomes for adult critical care admissions. The CMP ensures data accuracy through several steps, such as standardized dataset specifications, training courses, and both local and central data validation. [24] It applies established criteria from the Directory of Clinical Databases to assess data quality. several techniques used are data specification, validation, Evaluation Against Criteria that is CMP data was evaluated using criteria from the Directory of Clinical Databases, scoring a mean quality level of 3.4 out of 4. The results showed the CMPD contained validated data on over 129,000 admissions, with a variety of clinical metrics. The database is highly reliable and comparable to other UK critical care datasets.

In paper [25] author outlines the methodology and statistical analysis plan for creating and assessing clinical prediction models for pulmonary nodules. Where the limitations of existing models are discussed.

In paper [26] author develops a logistic regression diagnostic model to discriminate between benign and malignant solitary pulmonary nodules (SPNs). The model was constructed using clinical, biomarker, and radiological data from a training set of 1,679 patients (77.2% with malignant SPNs). [27] Independent variables included factors such as age, smoking history, family cancer history, nodule diameter, and CT characteristics (e.g., spiculation, lobulation, calcification). Techniques used include logistic regression for multivariable analysis and ROC (Receiver Operating Characteristic) curves to evaluate the model's performance in both training and test datasets. Additionally, the study compared this model with the Swensen model and the Li model, both of which also use logistic regression but with different independent factors and achieved high accuracy[28]

In paper [29] author discusses the development of the TRIPOD which stands for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis Initiative, which aims to improve the quality and transparency of reporting in studies related to prediction models. These models help healthcare providers estimate the risk of a disease (diagnostic models) or predict future events (prognostic models). [30] Despite their importance, prediction model studies often suffer from poor reporting, making it difficult to assess their bias and usefulness. The goal of TRIPOD is to enhance the standard and clarity of reporting in prediction model research, making it easier for readers and editors to evaluate the studies.

In paper [31] author describes the creation of the Lung Image Database Consortium that is LIDC and Image Database Resource Initiative (IDRI), which developed a publicly available repository of computed tomography (CT) scans for lung nodule detection and assessment. [32] This initiative involved collaboration among seven academic centres and eight medical imaging companies, resulting in a robust database of 1,018 clinical thoracic CT cases. Each case includes images and an XML file detailing a two-phase annotation process carried out by four thoracic radiologists are included with each case. In the initial phase, radiologists independently categorized lesions as "nodule greater than or equal to 3 mm," "nodule less than 3mm," or "non-nodule." In the second phase, they analysis their own marks alongside those of their peers to reach a final opinion. The database contains 7,371 lesions marked as nodules, with 2,669 classified as "nodule greater than or equal to 3 mm." This database aims to enhance medical imaging research related to lung nodules.

In paper [33] author talks through the usage of texture features in PET and CT imaging for cancer prognosis using benjamini-hochberg correction, optimum cut-off approach, kaplan-meier survival analysis concluding that the association between patient survival and texture features on PET or CT scans is not well supported by the available data.

In paper [34] author talks about using deep learning models like 17 3D convolutional neural network along with nodule detection to identify the abnormal regions within the lung scans. [35] The core of the system involves 17 different 3D CNN models, organized into two ensembles, which were combined for better performance, the models used Keras and Theano frameworks, Leaky ReLU activation functions and softplus regression layers, batch normalization, 3D convolution, and global max pooling layers, Extremely Randomized Trees (ERT) were used in a final classifier, which worked with nodule features predicted by the CNN models. The system also applied test-time augmentation. However the model was based on a limited dataset and a heavy computation resources were used. [36] The challenges involve dealing with high resolution medical images. In paper [20] Hugo J.W.L. et. al. Presents a radiomics approach to decode tumor phenotypes using non-invasive imaging, primarily focusing on lung and head-and-neck cancers, authors here used a quantitative radiomics method to extract 440 image features from CT scans. These features quantify tumor intensity, shape, texture, and wavelet decomposition where a multivariate Cox proportional hazards regression model was built using the top four radiomic features for prognosis. [37] Also taking us through existing technologies like RECIST and WHO measure tumor response to therapy using one- or two-dimensional size descriptors and TNM staging focuses on tumor resect ability. The developed model accomplished 0.65 for lung cancer data and 0.69 for head-and- neck cancer data.

In paper [38] author presents a radiomics approach to decode tumor phenotypes using non-invasive imaging, primarily focusing on lung and head-and-neck cancers, authors here used a quantitative radiomics method to extract 440 image features from CT scans. These features quantify tumor intensity, shape, texture, and wavelet decomposition where a multivariate Cox proportional hazards regression model was built using the top four radiomic features for prognosis. [39] Also taking us through existing technologies like RECIST and WHO measure tumor response to therapy using one- or two-dimensional size descriptors and TNM staging focuses on tumor resect ability. The developed model accomplished 0.65 for lung cancer data and 0.69 for head-and-neck cancer data.

In paper [40] author reviews the use of Multiple Imputation (MI) in 16 studies, highlighting variability in techniques and handling of missing data. Most studies used five or ten imputations, with no consensus on software. Rubin's rules were commonly used to combine estimates after MI, though some studies didn’t specify how this was done. Reported estimates included regression coefficients, hazard ratios, and confidence intervals from models like Cox regression, Poisson regression, and Weibull models. Some studies reported model performance but without detailing post-imputation calculations.

In paper [41] concludes, these findings indicate that the random assignment of heavy smokers to a smoking cessation program results in a significant reduction in the rate of pulmonary function decline, specifically a decrease of approximately 16%, corresponding to an improvement of over 10 ml/year in FEV1. Although we believe these results can be extrapolated to lighter smokers and potentially to female populations, validating this effect through randomized designs remains complex. Importantly, our data suggest that even among heavy and chronic smokers, cessation can yield beneficial effects, emphasizing that it is not too late to quit.

In paper [42] author takes us through risk assessment of malignancy in pulmonary nodules based on clinical and radiological factors, imaging techniques, follow-up protocols, and guidelines for surgical and non-surgical biopsy. Introduction of malignancy prediction calculators and improved algorithms to refine the risk assessment and management of pulmonary nodules. As well as provides detailed guidelines on the management of sub-solid nodules (SSNs).

In paper [43] author studies three pre-trained transfer learning models—EfficientNetB0, DenseNet169, and DenseNet201—were implemented to classify lung conditions, including COVID-19, Pneumonia, Tuberculosis, and healthy lungs. The dataset, consisting of 6,340 images, was split into training, testing, and validation sets in an 80:15:5 ratio. Image preprocessing techniques such as resizing, filtering, and augmentation were applied. Model performance was evaluated using confusion matrices and metrics like accuracy, precision, recall, F1-score, and rates such as TPR, TNR, FPR, and FNR to determine the optimal classification model. Young scientists can also learn from this paper how to develop CNN models that can be used to identify diseases early on in the process of using medical images.

In paper [44] author deliberates a follow-up study that found 2,545 men and 1,894 women died, with increasing mortality risk linked to lower FEV1 after adjusting for age, smoking, blood pressure, cholesterol, BMI, and social class. Relative hazard ratios for all-cause mortality in the lowest FEV1 quintile were 1.92 for men and 1.89 for women. Reduced FEV1 was also significantly associated with higher risk of death from ischemic heart disease, lung cancer, and stroke. For lifelong nonsmokers, reduced FEV1 was linked to all causes of death except cancer. Impaired lung function is a strong indicator of mortality risk, highlighting the importance of FEV1 in health assessments, especially for smokers. This study serves as a foundation for exploring the potential of using machine learning algorithms to improve the detection and prediction of lung diseases, aiming for more accurate and reliable outcomes.

DOI: 10.48175/IJARSCT-27926

III. PROPOSED ARCHITECTURE

The proposed framework is designed to accurately classify lung cancer subtypes—adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and normal—using a combination of advanced deep learning techniques. The methodology begins with data collection and preprocessing. Medical imaging datasets containing chest CT and X-ray scans are gathered from publicly accessible sources. These datasets are carefully curated to include balanced representations of each cancer subtype. All images are standardized by resizing them to a consistent dimension, typically 224×224 or 512×512 pixels, and normalized to a specific pixel intensity range, such as [0,1] or [-1,1]. Image augmentation techniques, including rotation, flipping, and scaling, are applied to improve model generalization and combat overfitting. The dataset is then split into training, validation, and testing sets to ensure comprehensive model evaluation.

To address the challenge of limited labeled data, Generative Adversarial Networks (GANs) are integrated into the workflow for unsupervised feature learning and synthetic image generation. The GAN consists of two main components: a generator and a discriminator. The generator creates synthetic lung cancer images that mimic real ones, while the discriminator attempts to distinguish between real and generated images. Through adversarial training, both networks improve iteratively, leading the generator to produce increasingly realistic images. These synthetic images not only enhance the diversity of the training dataset but also help the model learn underlying features more effectively, especially in cases where real labeled data is sparse.

For classification, a Convolutional Neural Network (CNN) is employed to categorize the images into one of the four lung cancer classes. The CNN architecture typically comprises multiple convolutional layers that extract spatial and hierarchical features from the input images. Pooling layers are used to downsample the feature maps, and fully connected layers are employed to interpret the learned features and make the final classification. The model is trained using categorical cross-entropy as the loss function and optimized using the Adam optimizer. To further enhance performance, pre-trained CNN architectures such as ResNet or DenseNet can be fine-tuned on the lung cancer dataset, leveraging transfer learning for better accuracy and faster convergence..

Model performance is evaluated using a range of metrics including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Cross-validation is performed to ensure the model's robustness and consistency across different data splits. This rigorous evaluation helps assess the model's ability to distinguish between the various lung cancer subtypes and identify potential areas for improvement.

Finally, the trained model is deployed through a user-friendly interface, either web-based or mobile, allowing healthcare professionals to upload CT or X-ray images and receive instant diagnostic predictions. This real-time inference capability can significantly assist in clinical decision-making. Future enhancements may involve integrating clinical metadata, applying the model to 3D CT volumes, and exploring federated learning approaches for decentralized training while maintaining patient data privacy.

1. Data Collection:

Datasets: The proposed framework utilizes publicly available datasets focused on lung cancer subtypes. These include:

a) CT scan datasets labeled for adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and normal cases

b) X-ray datasets identifying different lung cancer classes

c) Pediatric and adult chest imaging datasets containing various lung cancer categories

These datasets will be gathered and standardized through preprocessing steps that unify image formats and dimensions. The goal is to improve model generalization by applying techniques such as rotation, flipping, and zooming, which simulate real-world variations in medical imaging.

2. Data Preprocessing:

a) Resizing: All images will be resized to a uniform resolution (e.g., 224x224 pixels) to ensure compatibility with the CNN input layer.

b) Normalization: Pixel intensity values will be scaled to a common range (e.g., 0 to 1) for numerical stability during training.

DOI: 10.48175/IJARSCT-27926

c) Data Augmentation: Methods like flipping, rotating, scaling, and shifting will be applied to increase dataset diversity and minimize overfitting.

d) Image Splitting: The datasets will be divided into training, validation, and testing sets with balanced representation of all lung cancer subtypes to maintain fairness and prevent class imbalance.

III. SUPERVISED FEATURE LEARNING WITH CNNS

3.1 Convolutional Neural Networks (CNNs):

In this approach, a Convolutional Neural Network (CNN) is used to classify the lung cancer subtypes based on labeled CT scan and chest X-ray images. This model learns from the annotated data, extracting relevant features to distinguish between different subtypes of lung cancer, such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.

Feature Extraction: The CNN’s layers will learn to detect hierarchical patterns and features such as tumor shape, size, and texture, which are indicative of specific lung cancer subtypes.

Classification: After learning from the labeled data, the CNN will categorize the input images into one of the predefined lung cancer subtypes. This supervised approach ensures the model can make accurate predictions based on real-world clinical data.

3.2. Training the Model

The model is trained using both labeled data (for the CNN component) and unlabelled data (for the GAN component). The CNN is trained on labeled data where each chest X-ray or CT scan image is annotated with the correct lung cancer subtype (e.g., adenocarcinoma, squamous cell carcinoma, large cell carcinoma, or normal). The GAN, on the other hand, is trained on unlabelled data to generate realistic synthetic images that capture the underlying patterns and features of lung cancer images.

During training, the CNN will extract hierarchical features from labeled chest X-ray or CT scan images. It will learn to identify patterns and characteristics that distinguish between different lung cancer subtypes. The CNN’s weights are adjusted based on the ground truth labels in the dataset, refining its ability to classify images accurately.

The GAN component of the model is trained using unlabelled data, with the goal of learning to generate synthetic images that resemble real lung cancer images. The generator in the GAN creates synthetic images, while the discriminator attempts to distinguish between real and fake images. This process improves the generator’s ability to create realistic images, which will be used to augment the training dataset and improve the overall model's robustness.

4. Feature Extraction:

Once both the CNN and GAN are trained, the learned features from both components will be combined. The CNN will extract critical features for classification, while the GAN will provide additional synthetic images that help reinforce the learning process. These combined features will represent the most discriminative characteristics of the lung cancer subtypes and be used for classification tasks.

5. Disease Classification using CNN

After the CNN and GAN have learned the necessary features, the CNN will be used to classify the lung cancer subtypes. It will classify new, unseen chest X-ray or CT scan images into one of the cancer subtypes based on the learned features.

CNN Architecture:

The CNN will consist of several convolutional layers that progressively capture higher-level features, followed by pooling layers for dimensionality reduction. The fully connected layers at the end of the network will use the extracted features to classify the images into one of the subtypes.

DOI: 10.48175/IJARSCT-27926

Loss Function and Optimization:

For the classification task, cross-entropy loss will be used, and the model will be optimized using the Adam optimizer. This process minimizes the loss function, improving the accuracy of the model's predictions for the lung cancer subtypes.

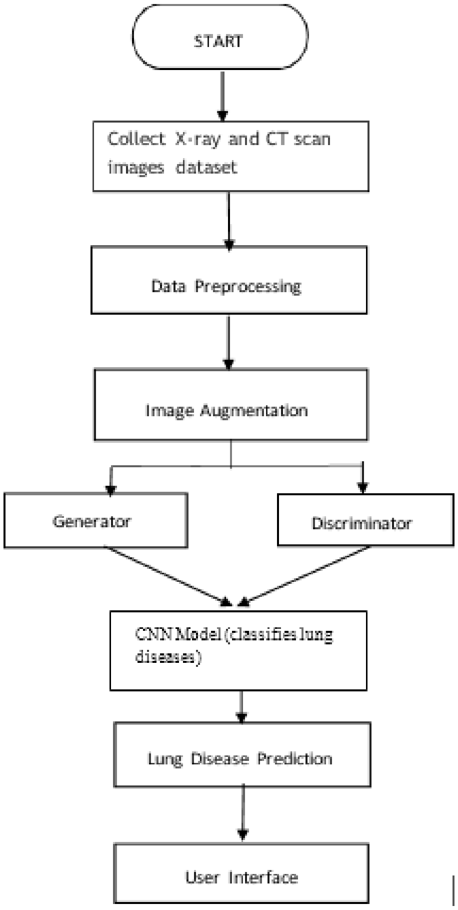


Fig. 1. Architecture and Flow of System

IV. EXPERIMENTAL SETUP

To conduct the experiment, specific minimum hardware requirements must be met, including a solid-state drive (SSD) with a capacity of 512 GB for storage, CPU that is central processing unit accompanied by a graphics processing unit (GPU), and a minimum memory (RAM) of 16 GB, with 32 GB recommended for optimal performance. Additionally, an adequate cooling system is necessary to maintain optimal operating temperatures. The software requirements consist of a Windows operating system, a development environment utilizing Jupyter Notebook, Visual studio and Python version 3.10.11, and essential machine learning libraries and frameworks such as TensorFlow, scikit-learn, pandas, and OpenCV. For model deployment, frameworks like Flask or Django should be employed to facilitate application deployment.

V. RESULT AND ANALYSIS OF ALGORITHMS

The heading of the Acknowledgment section and the References section must not be numbered. Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template.

	Bagging	Logistic Regression	GAN + CNN
Accuracy	88.60%	88%	95%
Precision	~0.85	~0.87	~0.92
F1 score	~0.58	~0.56	~65
Recall	~0.86	~0.88	~90

Accuracy: It is the percentage of accurate predictions true positives and true negatives that are made relative to all of the forecasts that were made. It indicates how frequently the model is accurate overall.

Accuracy= True Positives + True Negatives Total Predictions

Precision: It measures the accuracy of positive predictions how many of the predicted positives are actually correct.

Precision= True Positives + False Positives True Positives

Recall: It gauges how well a model can detect every positive case and is also referred to as sensitivity or true positive rate. It provides an answer to the following query: how many real benefits did the model accurately identify.

Recall= True Positives + False Negatives True Positives

The performance comparison in Table 1 shows that the CNN+GAN model outperforms Bagging and Logistic Regression (LR) across all metrics. CNN+GAN achieves the highest accuracy (95%), precision (~0.92), recall (~0.90), and F1 score (~0.65), indicating superior overall performance. Bagging and LR show similar accuracy (~88%) but lower precision and F1 scores. The results highlight the effectiveness of combining CNN with GAN for improved classification accuracy and balanced prediction.

This section presents the visual and quantitative results of the proposed model/method. The input images are processed through the system, and the output images/results demonstrate the effectiveness of the approach.

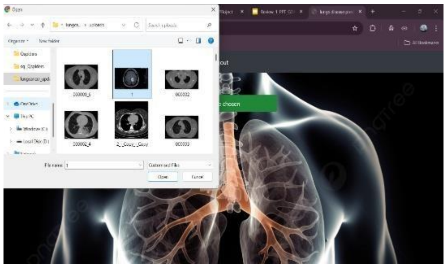
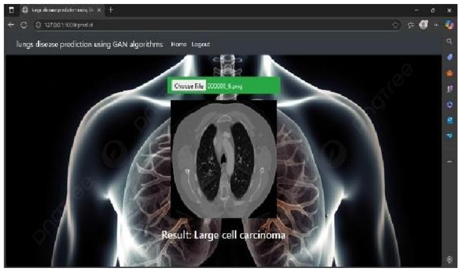


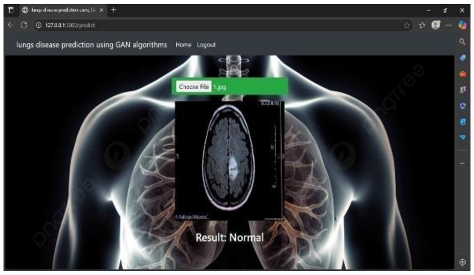
Fig.2. Taking an Input Image


127.0.0.1:5000/predict

lungs disease prediction using GAN algorithms

Result: Large cell carcinoma

Fig.3. Predicting Disease



Result: Normal

Fig.4. Predicting Normal

VI. CONCLUSION

In this study, we developed a deep learning-based framework named Lung-GAN for the accurate prediction of lung diseases using chest CT and X-ray images. By integrating Convolutional Neural Networks (CNNs) with Generative Adversarial Networks (GANs), we addressed the challenge of limited annotated medical data through effective synthetic data augmentation. Our model was trained to classify four distinct categories: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal lung tissue.

The performance analysis demonstrated that our CNN+GAN approach outperformed traditional machine learning methods such as Bagging, Logistic Regression (LR), and Random Forest in all evaluation metrics. It achieved the highest accuracy (95%), precision (\~0.92), recall (\~0.90), and F1 score (\~0.65), indicating improved classification capability and robustness even with constrained data availability.

These promising results highlight the potential of the proposed Lung-GAN framework to aid in early diagnosis, reduce radiologist workload, and enhance clinical decision-making in lung cancer treatment. Future work may focus on integrating more diverse datasets and expanding the model's capability to detect other pulmonary conditions.

REFERENCES

[1]. Vishal Borate, Dr. Alpana Adsul, Palak Purohit, Rucha Sambare, Samiksha Yadav, Arya Zunjarrao, "A Role of Machine Learning Algorithms for Lung Disease Prediction and Analysis," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 4, Issue 3, pp. 425-434, October 2024, DOI: 10.48175/IJARSCT-19962.

[2]. V. K. Borate and S. Giri, "XML Duplicate Detection with Improved network pruning algorithm," 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 2015, pp. 1-5, doi: 10.1109/PERVASIVE.2015.7087007.

[3]. Borate, Vishal, Alpana Adsul, Aditya Gaikwad, Akash Mhetre, and Siddhesh Dicholkar. "Analysis of Malware Detection Using Various Machine Learning Approach," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 4, Issue 2, pp. 314-321, November 2024, DOI: 10.48175/IJARSCT-22159.

[4]. Borate, Mr Vishal, Alpana Adsul, Mr Rohit Dhakane, Mr Shahuraj Gawade, Ms Shubhangi Ghodake, and Mr Pranit Jadhav. "A Comprehensive Review of Phishing Attack Detection Using Machine Learning Techniques," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 4, Issue 2, pp. 435-441, October 2024 DOI: 10.48175/IJARSCT-19963.

[5]. Akanksha A Kadam, Mrudula G Godbole, Vaibhavi S Divekar, Vishakha T. Mandage and Prof. Vishal K Borate, "FIRE ALARM AND RESCUE SYSTEM USING IOT AND ANDROID", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.11, Issue 2, Page No pp.815-821, May 2024.

[6]. Prof. Vishal Borate, Prof. Aaradana Pawale, Ashwini Kotagonde,Sandip Godase and Rutuja Gangavne, "Design of low-cost Wireless Noise Monitoring Sensor Unit based on IOT Concept", International Journal of Emerging

Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.10, Issue 12, page no.a153-a158, December-2023.

[7]. Dnyanesh S. Gaikwad, Vishal Borate, "A REVIEW OF DIFFERENT CROP HEALTH MONITORING AND DISEASE DETECTION TECHNIQUES IN AGRICULTURE", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.10, Issue 4, Page No pp.114-117, November 2023.

[8]. Prof. Vishal Borate, Vaishnavi Kulkarni and Siddhi Vidhate, "A Novel Approach for Filtration of Spam using NLP", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.10, Issue 4, Page No pp.147-151, November 2023.

[9]. Prof. Vishal Borate, Kajal Ghadage and Aditi Pawar, "Survey of Spam Comments Identification using NLP Techniques", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.10, Issue 4, Page No pp.136-140, November 2023.

[10]. Akanksha A Kadam, Mrudula G Godbole, Vaibhavi S Divekar and Prof. Vishal K Borate, "Fire Evacuation System Using IOT & AI", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.10, Issue 4, Page No pp.176-180, November 2023.

[11]. Shikha Kushwaha, Sahil Dhankhar, Shailendra Singh and Mr. Vishal Kisan Borate, "IOT Based Smart Electric Meter", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 3, pp.51-56, May-June-2021.

[12]. Nikita Ingale, Tushar Anand Jha, Ritin Dixit and Mr Vishal Kisan Borate, “College Enquiry Chatbot Using Rasa,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 3, pp.201-206, May-June-2021.

[13]. Pratik Laxman Trimbake, Swapnali Sampat Kamble, Rakshanda Bharat Kapoor, Mr Vishal Kisan Borate and Mr Prashant Laxmanrao Mandale, "Automatic Answer Sheet Checker,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 3, pp.212-215, May-June-2021.

[14]. Shikha Kushwaha, Sahil Dhankhar, Shailendra Singh and Mr. Vishal Kisan Borate, "IOT Based Smart Electric Meter"” International Journal of Scientific Research in Science and Technology (IJSRST), ISSN: 2395-602X, Volume 5, Issue 8, pp.80-84, December-2020.

[15]. Nikita Ingale, Tushar Anand Jha, Ritin Dixit and Mr Vishal Kisan Borate, “College Enquiry Chatbot Using Rasa,” International Journal of Scientific Research in Science and Technology (IJSRST), ISSN: 2395-602X, Volume 5, Issue 8, pp.210-215, December-2020.

[16]. Pratik Laxman Trimbake, Swapnali Sampat Kamble, Rakshanda Bharat Kapoor and Mr Vishal Kisan Borate, “Automatic Answer Sheet Checker,” International Journal of Scientific Research in Science and Technology (IJSRST), ISSN: 2395-602X, Volume 5, Issue 8, pp.221-226, December-2020.

[17]. Chame Akash Babasaheb, Mene Ankit Madhav, Shinde Hrushikesh Ramdas, Wadagave Swapnil Sunil, Prof. Vishal Kisan Borate, " IoT Based Women Safety Device using Android, International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 5, Issue 10, pp.153-158, March-April-2020.

[18]. Harshala R. Yevlekar, Pratik B. Deore, Priyanka S. Patil, Rutuja R. Khandebharad, Prof. Vishal Kisan Borate, " Smart and Integrated Crop Disease Identification System, International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 5, Issue 10, pp.189-193, March-April-2020.

[19]. Yash Patil, Mihir Paun, Deep Paun, Karunesh Singh, Vishal Kisan Borate, " Virtual Painting with Opencv Using Python, International Journal of Scientific Research in Science and Technology(IJSRST), Online ISSN : 2395- 602X, Print ISSN : 2395-6011, Volume 5, Issue 8, pp.189-194, November-December-2020.

[20]. Mayur Mahadev Sawant, Yogesh Nagargoje, Darshan Bora, Shrinivas Shelke and Vishal Borate, Keystroke Dynamics: Review Paper International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 10, October 2013.

DOI: 10.48175/IJARSCT-27926

[21]. Modi, S., Sale, D., Borate, V., Mali, Y.K. (2025). Enhancing Learning Outcomes Through the Use of Conducive Learning Spaces. In: Majumder, M., Zaman, J.K.M.S.U., Ghosh, M., Chakraborty, S. (eds) Computational Technologies and Electronics. ICCTE 2023. Communications in Computer and Information Science, vol 2376. Springer, Cham. https://doi.org/10.1007/978-3-031-81935-3_4.

[22]. Y. Mali, M. E. Pawar, A. More, S. Shinde, V. Borate and R. Shirbhate, "Improved Pin Entry Method to Prevent Shoulder Surfing Attacks," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10306875.

[23]. Nadaf, N., Waghodekar, P., Magdum, A., Gupta, P., Borate, V.K., Mali, Y.K. (2025). Architecture for Cost-Effective Deployment of Models to Transfer Style Across Images. In: Shukla, P.K., Bhatt, A., Mittal, H., Engelbrecht, A. (eds) Computer Vision and Robotics. CVR 2024. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-97-8868-2_44.

[24]. Modi, S., Mali, Y., Sharma, L., Khairnar, P., Gaikwad, D.S., Borate, V. (2024). A Protection Approach for Coal Miners Safety Helmet Using IoT. In: Jain, S., Mihindukulasooriya, N., Janev, V., Shimizu, C.M. (eds) Semantic Intelligence. ISIC 2023. Lecture Notes in Electrical Engineering, vol 1258. Springer, Singapore. https://doi.org/10.1007/978-981-97-7356-5_30

[25]. Waghodekar, P. et al. (2025). Security Protecting Confirmation of IoMT in Distributed Cloud Computing. In: Shukla, P.K., Bhatt, A., Mittal, H., Engelbrecht, A. (eds) Computer Vision and Robotics. CVR 2024. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-97-8868-2_3

[26]. Rojas, Macedo, and Yolaina Malí. "Programa de sensibilización sobre norma técnica de salud N° 096 MINSA/DIGESA V. 01 para la mejora del manejo de residuos sólidos hospitalarios en el Centro de Salud Palmira, Independencia-Huaraz, 2017." (2017).

[27]. Sale, D., Khare, N., Kadam, S., Mali, Y.K., Borate, V., Gaur, A. (2025). A Secure Pin Entry Mechanism for Online Banking by Defending Shoulder-Surfing Attacks. In: Kumar, S., Mary Anita, E.A., Kim, J.H., Nagar, A. (eds) Fifth Congress on Intelligent Systems. CIS 2024. Lecture Notes in Networks and Systems, vol 1278. Springer, Singapore. https://doi.org/10.1007/978-981-96-2703-5_4

[28]. Rathod, V.U., Nandgoankar, V., Dhawas, N., Mali, Y.K., Chaudhari, H., Patil, D. (2025). Smart Traffic Light Management System Using IoT and Deep Learning. In: Singh, S., Arya, K.V., Rodriguez, C.R., Mulani, A.O. (eds) Emerging Trends in Artificial Intelligence, Data Science and Signal Processing. AIDSP 2023. Communications in Computer and Information Science, vol 2439. Springer, Cham. https://doi.org/10.1007/978-3-031-88759-8_9.

[29]. Kale, Hrushikesh, Kartik Aswar, and Dr Yogesh Mali Kisan Yadav. "Attendance Marking using Face Detection." International Journal of Advanced Research in Science, Communication and Technology: 417-424.

[30]. Inamdar, Faizan, Dev Ojha, C. J. Ojha, and D. Y. Mali. "Job Title Predictor System." International Journal of Advanced Research in Science, Communication and Technology (2024): 457-463.

[31]. Jagdale, Sudarshan, Piyush Takale, Pranav Lonari, Shraddha Khandre, and Yogesh Mali. "Crime Awareness and Registration System." International Journal of Scientific Research in Science and Technology 5, no. 8 (2020).

[32]. Suoyi, Han, Yang Mali, Chen Yuandong, Yu Jingjing, Zhao Tuanjie, Gai Junyi, and Yu Deyue. "Construction of mutant library for soybean'Nannong 94-16'and analysis of some characters." Acta Agriculturae Nucleatae Sinica 22 (2008).

[33]. Van Wyk, Eric, and Yogesh Mali. "Adding dimension analysis to java as a composable language extension." In International Summer School on Generative and Transformational Techniques in Software Engineering, pp. 442-456. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[34]. Mali, Yogesh, Vijay U. Rathod, Ravindra S. Tambe, Radha Shirbhate, Deepika Ajalkar, and Priti Sathawane. "Group-Based Framework for Large Files Downloading." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-4. IEEE, 2023.

[35]. Modi, Shabina, Deepali Sale, Vishal Borate, and Yogesh Kisan Mali. "Enhancing learning outcomes through the use of conducive learning spaces." In International Conference on Computational Technologies and Electronics, pp. 45-53. Cham: Springer Nature Switzerland, 2023.

DOI: 10.48175/IJARSCT-27926

ISSN
2581-9429
IJARSCT

[36]. Mali, Yash, Himani Malani, Nishad Mahore, and Rushikesh Mali. "Hand Gesture Controlled Mouse." International Research Journal of Engineering and Technology (2022).

[37]. Gai Mali, Yustinus Calvin. "The exploration of Indonesian students' attributions in EFL reading and writing classes." Bahasa dan Seni: Jurnal Bahasa, Sastra, Seni, dan Pengajarannya 50, no. 1 (2022): 1.

[38]. Mali, Y. "Effort attributions in Indonesian EFL classrooms." Jurnal Ilmu Pendidikan 22, no. 1 (2016): 80-93.

[39]. Malî, Yôsef, ed. Narrative patterns in scientific disciplines. Cambridge University Press, 1994.

[40]. Das et al., "Antibiotic susceptibility profiling of Pseudomonas aeruginosa in nosocomial infection," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10723982.

[41]. Dhokale, Bhalchandra D., and Ramesh Y. Mali. "A Robust Image Watermarking Scheme Invariant to Rotation, Scaling and Translation Attack using DFT." International Journal of Engineering and Advanced Technology 3, no. 5 (2014): 269.

[42]. Yogesh Mali, NilaySawant, "Smart Helmet for Coal Mining," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)Volume 3, Issue 1, February 2023,DOI: 10.48175/IJARSCT-8064

[43]. Mali, Yash, Anuja Tambade, Mrunmayi Magdum, and B. G. Patil. "Artificial Neural Network Based Automatic Number Plate Recognition System." International Journal on Recent and Innovation Trends in Computing and Communication 4, no. 5 (2016): 128-131.

[44]. Mali, Y., and E. Deore. "Design and Analysis with Weight Optimization of Two Wheeler Gear Set." International Advanced Research journal in Science, Engineering and Technology 4, no. 7 (2017).

DOI: 10.48175/IJARSCT-27926