Role of Automated Machine Learning approach for lung cancer data classification using modern techniques

Dr Saurabh Mishra[1*], Dr. Rishabh Chitranshi[2], Dr Shashwat Katiyar[3], Dr. Amit Kumar Jana[4], Dr. Suprakash Pradhan[5]

[*1]*Assistant professor, Department of Biotechnology and Life sciences mangalayatan University Aligarh, Email: mishra.saurabh228@gmail.com*
[2]*Chitranshi, School of Biological Engineering and Sciences, Shobhit University, Gangoh, Saharanpur, UP-247341*
[3]*Assistant Professor, Department of Biochemistry, School of Life Sciences and Biotechnology, Chhatrapati Shahu Ji Maharaj University, Kanpur (UP), India, Email: shashwatkatiyar8@gmail.com*
[4]*State Aided College Teacher -1, College: Debra Thana Sahid Kshudiram Smriti Mahavidyalaya (Autonomous), Department : Nutrition (UG&PG), Email: amitnutrition@gmail.com*
[5]*State Aided College T-1, Dept of Nutrition, Prabhat Kumar College, Contai, East Medinipur, West Bengal, Email: pradhansuprakash@gmail.com*

Corresponding Author
Dr Saurabh Mishra
Assistant professor, Department of Biotechnology and Life sciences mangalayatan University Aligarh, Email: mishra.saurabh228@gmail.com

ABSTRACT
Medical data classification is the process for categorizing healthcare-related information, such as patient records, medical images, and diagnostic results, into distinct groups based on predefined criteria. Machine learning plays a pivotal role by classifying the medical data. Machine learning and deep learning techniques have become indispensable tools in precision healthcare and automated disease diagnosis, enabling accurate clinical decision-making and predictive analytics [22,23,27]. leveraging algorithms that can detect pattern automatically from the data and can make predictions with less human interference. ML models unlike traditional method analyze vast amount of medical data, recognize complex patterns , and generalize new cases according to learning., This paper therefore seeks to produce informed and precise medical data classification strategy for Lung cancer datasets. For lung cancer data classification Machine learning algorithms like Support vector machine(SVM), Convolutional Neural Networks(CNN), Residual Neural Network(ResNet), and Random Forest(RF) are used in this research paper due to their ability to extract meaningful features from medical images and clinical data. CNN are specifically effective in medical images, making them ideal for detecting lung nodules and classify them as benign or malingnant. ResNet ,which deep learning further enhances classification accuracy by using convolutional layers. SVM is widely used for classification task due to its effectiveness in handling small datasets and high-dimensional spaces.

Among them, ResNet50 achieved the highest accuracy of 91.3%, making it the most effective for multi-class classification (Normal, Benign, Malignant). CNN followed with 85.6%, while RF and SVM achieved 84.4% and 83.2% respectively.

Keywords—Lung Cancer, Data Classification, IQ-OTH/NCCD Dataset, Convolutional Neural Network, Deep Learning, Medical Imaging, Benign ,Malignant Tumors.

How to cite this article: Mishra S, Chitranshi R, Katiyar S, Jana AK, Pradhan S. Role of Automated Machine Learning Approach for Lung Cancer Data Classification Using Modern Techniques. Int J Drug Deliv Technol. 2026;16(72s): 744-756. DOI: 10.25258/ijddt.16.72s.82

INTRODUCTION

Throughout the world, "Lung Cancer" is the primary reason of cancer mortality. [1]Lung cancer is one of the most prevalent deadly forms of cancer with early detection ,diagnosis plays important role in improving patient survival rates, as timely intervention can lead to more effective treatment and better prognoses. One of the reasons lung cancer ranks first on the list is that it is frequently not discovered until the disease is advanced. [2,21] Early categorization of the help the treatment of certain cancers run more smoothly and permit's just say rescue millions of people each year. The and accurate classification of lung cancer types is the motivation.[2] Research has shown that early discovery relates to decreased lung cancer fatalities. [4] Despite declining Lung cancer rates and fatalities remain still quite high despite years of attempts to lower smoking and encourage stopping because of the rising risk of the condition among older people. From 2015 to 2065, Jeon et al. estimate a 79% age-adjusted mortality drop thanks to tobacco restriction initiatives; however, during this time period 4.4 million lung cancer fatalities are expected. [5]

The human mind is finite in its capability to manage and completely use the gathering of so much data. Integrating and evaluating these vast and complicated datasets, which have comprehensively defined lung cancer by means of several angles, depends critically on machine learning-based methods. [6] Among other fields, ML finds application in military, finance, and medical. Currently, a large quantity of data is easily obtainable. ML, which is a technically advanced and innovative field, has become incredibly popular in the industry over time.[7] Therefore, it is essential to analyze this data in order to extract any useful information. The use of machine learning algorithms in image recognition, speech recognition and automation is a new trend, as they are both cheap and precise..[8] Medical data categorization is viewed as a difficult task in the area of medical informatics.[9] Recently, machine learning algorithms have been utilized to detect trends in medical data and have demonstrated remarkable accuracy in forecasting these patterns. ML methods generate accurate predictions influenced by previous experiences. As a result, it serves as an excellent strategy that assists in disease forecasting and identification. [10,28] This study focuses on three types of lung cancer datasets normal, malignant, and Benign from IQ-OTH/NCCD-Lung cancer Dataset, which consists of CT scan images of lung cancer patients. The classification involves various machine learning algorithms including Support vector machine(SVM), Convolutional Neural Networks(CNN), Residual Neural Network(ResNet), Visual Geometry Group(VGG), and Random Forest(RF).

The models are assessed based on its accuracy, precision, recall, F1-score, and area under the curve. This study ensures better data generalization , therefore improving diagnostic accuracy and reducing human errors false positives and false negatives. ML techniques make correct predictions based on the past experiences. Therefore, it is a great approach which aids in the disease prediction and diagnosis.[10]

LITERATURE REVIEW

In 2021 , Sultana Abida, Khan Tahsin Tasnia and Hossian Tahmin [3] classifies early lung cancer types ehich is crucial for effective treatment and saving lives. It focuses on automated classification of datasets that consists 15,000 CT scan images categorized into three types -Lung benign tissue, Lung adenocarcinoma, and Lung squamous cell carcinoma.

Five deep learning models are 2D CNN with SVM , ResNet-50, InceptionResNetV2, Inception-V3 AND VGG-19 –were evaluated on basis of their accuracy, precision ,recall, and F1 score. In results Inception-V3 achieved the highest validation accuracy of 99.13% making is effective for classification of lung cancer data.[3] Recent studies have also demonstrated the effectiveness of transfer learning architectures for multiclass medical image classification, achieving high diagnostic accuracy in tumor detection tasks [23,24].

In 2021, Ali Imdad, Muzammil Muhammad, Haq Ihsan UI, Amir Muhamad, Adullah Suheel [10]enhances their study on Computer-Aided Diagnosis(CAD) for lung cancer proposing
decision-level fusion technique for lung nodule classification in CT images. The system utilizes SVM and AdaBoostM2 classifiers along with deep features from a range of architectures, such as VGG-16, VGG-19, GoogLeNet, Inception-V3, ResNet-18, ResNet-50, ResNet-101, and InceptionResNet-V2.The classification accuracy by identifying optimal feature extraction layers accuracy increased fro 76.88% to 86.28% (ResNet-101) and from 67.37% to 83.40% (GoogLeNet) while reducing the errors. The findings SVM to be more effective than AdaBoostM2, achieving 90[10].46% accuracy on Lungx datasets.

In 2022, Naresh Kumar, Manoj Sharma, Vijay Pal Singh Charanjeet Madan, Seema Mehandia [4] The study presents six techniques for extracting handcrafted features based on color, texture, shape, and structure. Handcrafted attributes from Gradient Boosting (GB), SVM-RBF, Multilayer Perceptron (MLP), and Random Forest (RF) classifiers are developed and evaluated for the classification of lung and colon cancer. [11] CNN networks have demonstrated outstanding performance regarding recall, accuracy, precision, F1 score, and ROC-AUC. The RF classifier utilizing DenseNet-121 to extract deep features is capable of accurately identifying lung and colon cancer tissue with a recall and accuracy of 98. 60%.[11]

In 2023, Li Yuting, Yan Bingshou, He shiming[12.29]studied on Lung cancer accounts that proportion of malingnant tumors are relatively high.As they represent the most common form of non-small cell lung cancer (NSCLC), they are associated with high rates of mortality and morbidity. The management of cancer includes surgical removal, chemotherapy, targeted treatment, and radiation therapy; however, these approaches have a low 5-year survival rate. The article explores cutting-edge developments in nano drug delivery systems, molecular targeted therapies, photothermal treatment methods, and immunotherapy for lung cancer. [5]The objective is to improve treatment results and create a theoretical framework for the creation of new lung cancer medications.

In 2024 , Jalaja, S and Prerita, H G [13]implemented automated detection method to improve accuracy and manual efforts. The research integrates ML (KNN, SVM) and deep learning (CNN, VGG16 transfer learning) and feature extraction techniques like PCA and HOG to classify lung cancer data efficiently. The paper evaluates models performance, demonstrating that CNN achieves highest accuracy(96%-99%).The findings contribute in detection of early lung cancer, improving patient care. Similar hybrid architectures combining VGG-based feature extraction with SVM classifiers have shown promising results in chest image classification applications [25].

In 2024, Diwedi Himanshu Kumar , Misra Anuradha Tiwari and Amod Kumar [14]proposed an Enhanced CNN with progressive Transfer Learning (ECNNPTL), Integrating a modified ResNet50 for extraction of features and an optimized support vector Machine (OVSM) for classification. The models have two-stage training process, where pre-trained ResNet50 layers are progressively trained with increasing image resizes (64 to 256) pixels. The results demonstrate that ResNet50+OVSM model achieves 96.8% accuracy in testing and 98.5% accuracy in training, making highly effective approach for medical classification

In 2024, Vemula Suseela Triveni, Sreevani Maddukuri, Rajarjeswari Perepi, Bhargavi Kumbham, Tavares Joao Manuel R.S, Alankrita Sampath[15].This document utilizes CT scan datasets to train a machine learning model, Convolutional Neural Networks (CNN), to autonomously learn and adjust features in lung images, enhancing classification precision and lung cancer diagnostic precision.. It paper provides a comparative analysis of the performance of CNN, VGG-16, and VGG-19. Notably, the built transfer learning model VGG-16 achieved a remarkable accuracy of 95%, surpassing the baseline method.[15]

METHADOLOGY

Objectives of the study

The primary objective of this study is to develop and evaluate machine learning and deep learning models for the accurate classification of lung cancer utilizing CT scan images. The research seeks to accomplish the following specific objectives:

- To automate the categorization of lung cancer into normal, benign, and malignant classifications by utilizing image-based machine learning techniques.
- To evaluate the effectiveness of conventional machine learning algorithms (e. g. , Support Vector Machine, Random Forest) against deep learning frameworks (e. g. , CNN, ResNet, VGG).
- To determine the most efficient model regarding accuracy, precision, recall, F1-score, and AUC-ROC, for dependable lung cancer detection.
- To implement preprocessing and feature extraction techniques that improve model performance and generalizability.
- To diminish human error and shorten diagnostic time by offering a rapid and precise automated solution for medical professionals

These objectives direct the comprehensive methodology process, spanning from dataset preparation to model assessment.

Dataset

This study employs the IQ-OTH/NCCD Lung Cancer Dataset, a publicly available medical imaging dataset specifically designed for the purpose of classifying lung cancer through the utilization of CT scans. The dataset consists of a total of 1,190 CT scan images, each categorized into one of three unique classifications: Normal, signifying the absence of lung cancer; Benign, denoting non-cancerous irregularities; and Malignant, pertaining to cancerous lung tumors.[16]

These images are derived from authentic clinical cases, rendering the dataset not only varied but also applicable for the advancement of diagnostic systems. The equitable distribution of images among all three categories guarantees impartial and unbiased model training, which is essential for attaining consistent classification accuracy across various lung conditions.[16]

To prepare the data for machine learning and deep learning models, a succession of preprocessing steps was implemented. All images were resized to 224x224 pixels to ensure consistent input dimensions, especially for compatibility with convolutional neural networks such as VGG and ResNet. Normalization was implemented to adjust pixel values within the range of 0 to 1, thereby enhancing the training speed and convergence of the models. Furthermore, data augmentation techniques, including rotation, horizontal and vertical flipping, zooming, and shifting, were utilized to augment the variability and robustness of the dataset. Advanced preprocessing and augmentation strategies have been reported to significantly improve the robustness and classification accuracy of deep learning models in medical image analysis [25]. This aids in simulating actual conditions and diminishes the risk of overfitting, particularly in deep learning frameworks. The categorical labels were transformed through one-hot encoding for deep learning models and label encoding for classical machine learning algorithms. This meticulously organized and preprocessed dataset provides a robust basis for the construction, training, and assessment of various AI models aimed at detecting lung cancer.[16]

Data Interpretation Techniques

To guarantee significant and dependable insights from the lung cancer classification models, various statistical and analytical methodologies were utilized for the purpose of interpreting the results:

- Confusion Matrix

Confusion matrices were constructed to illustrate accurate and inaccurate predictions across the three categories: Normal, Benign, and Malignant. This facilitated the identification of the model's strengths and weaknesses in class-specific predictions.

- Classification Report

Each model was assessed utilizing precision, recall, and F1-score, offering a comprehensive perspective on performance that extends beyond mere accuracy. Metrics were averaged employing macro and weighted approaches to address class imbalance.

- Accuracy Score

Overall accuracy was utilized as a fundamental metric. Although straightforward to interpret, it was supplemented by additional metrics to guarantee a dependable evaluation.

- ROC-AUC Curve (for Binary Cases)

For binary assessments (e. g. , Malignant vs. Non-Malignant), ROC curves and AUC scores were illustrated to comprehend the model’s capacity for discrimination.

- Visualization Tools

Graphs depicting the comparison between training and validation loss, as well as accuracy, in conjunction with bar plots and heatmaps, were utilized to effectively convey performance comparisons. PCA scatter plots facilitated the visualization of feature separability in diminished dimensions.

Model Development – ResNet50

One of the principal models utilized in this research is ResNet50, a deep convolutional neural network that has demonstrated outstanding performance in medical image classification assignments. ResNet, also known as Residual Network, addresses the issue of vanishing gradients in deep architectures by incorporating skip connections, thereby facilitating a more effective flow of gradients through the network during backpropagation.[17]

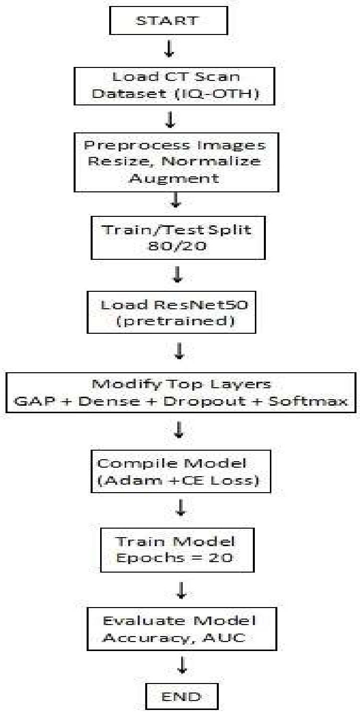
For this study, the pretrained ResNet50 architecture was employed through transfer learning, in which the foundational convolutional layers of the model were preserved (pretrained on ImageNet), and the upper classification layers were substituted with customized dense layers conducive to categorizing the CT scan images into three classifications: Normal, Benign, and Malignant.[17]

- Implementation Details

o Input Size: All images were resized to 224x224 pixels with three channels.
o Architecture Modification: The upper layer of ResNet50 was eliminated and substituted with:
o A Global Average Pooling layer
o A fully connected Dense layer comprising 256 units and utilizing ReLU activation
o A Dropout layer (rate = 0. 3) to mitigate overfitting
o An output Dense layer with three units and softmax activation for multiclass classification
o Optimizer: Adam
o Loss Function: Categorical Crossentropy
o Batch Size: 32
o Epochs: 20

o Validation Split: 20% of training data

The model was assembled and trained utilizing the Keras deep learning framework. Data was randomized and augmented through ImageDataGenerator to introduce variability and enhance the model's generalization performance.[17]



The model attained a high level of accuracy throughout the training and validation stages, with loss metrics demonstrating consistent convergence. This affirms the ability of deep residual networks to acquire complex patterns from CT scans and to differentiate between various lung conditions effectively.[17]

Model Development – Convolutional Neural Network (CNN)

In conjunction with ResNet50, a tailored Convolutional Neural Network (CNN) was devised and trained from the ground up to categorize lung CT scans into Normal, Benign, and Malignant classifications. CNNs are highly effective for image analysis owing to their capacity to autonomously acquire spatial hierarchies of features.[17]

Model Architecture:

The CNN model adheres to a structured architecture intended to systematically extract features and execute classification:

o Input Layer: Accepts images resized to 224x224 pixels, comprising 3 color channels (RGB).

o Convolutional Layers: Consists of three blocks of convolution layers utilizing ReLU activation, each succeeded by MaxPooling2D to diminish spatial dimensions.

o Dropout Layers: Implemented subsequent to convolution blocks (Dropout = 0. 25) to mitigate overfitting.

o Flatten Layer: Converts the feature map into a 1D vector.

o Fully Connected Dense Layers: Comprises two Dense layers featuring ReLU activation (e. g. , 128 and 64 neurons).

o Output Layer: A Dense layer containing 3 neurons and softmax activation for multiclass classification.

Training Configuration:

o Optimizer: Adam

o Loss Function: Categorical Crossentropy

o Batch Size: 32

o Epochs: 25

o Validation Split: 20%

• Image Augmentation & Preprocessing:

The model was trained utilizing an augmented dataset generated with ImageDataGenerator, implementing:

o Rotation

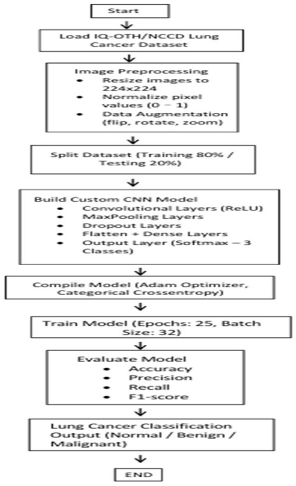
o Zoom

o Horizontal flip

o Width and height shifts

All images were standardized to a pixel range of 0–1 to enhance convergence.

The model exhibited consistent training and validation accuracy trends, signifying effective learning from CT scan images. The CNN has been found to be advantageous as a lightweight baseline when compared to more intricate architectures such as ResNet50.[17]



Model Development – Support Vector Machine (SVM)

To enhance deep learning models, a Support Vector Machine (SVM) was employed as a traditional machine learning benchmark. SVMs are well-recognized for their efficiency in binary and multiclass classification challenges, particularly in high-dimensional environments, thereby rendering them appropriate for medical image-based predictive tasks such as lung cancer classification.[17]

Feature Extraction

Since Support Vector Machines (SVMs) function on structured data instead of unprocessed images, the CT scan images from the IQ-OTH/NCCD dataset were initially processed through a feature extraction pipeline. A pretrained Convolutional Neural Network (CNN) (such as VGG16 or a bespoke shallow CNN) was utilized to extract deep features from the intermediate layers, transforming each image into a high-dimensional feature vector.[17]

SVM configuration

o Kernel: Radial Basis Function (RBF)

o Feature Vector: Extracted utilizing CNN (e. g. , from the flatten or dense layer)

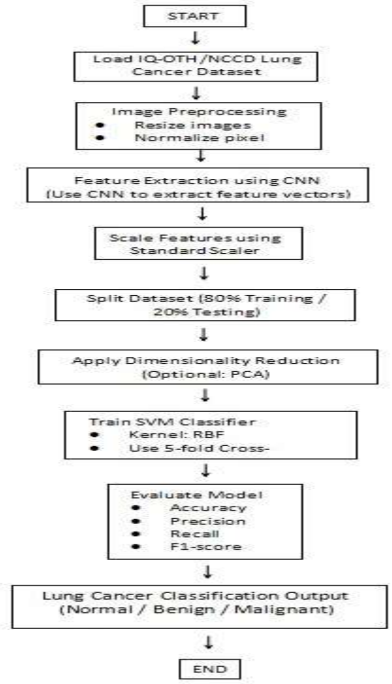
o Scaler: StandardScaler was utilized to normalize the feature vectors

o Train-Test Split: 80% allocated for training and 20% allocated for testing

o Cross-validation: 5-fold cross-validation was implemented to guarantee robustness.

Training and Evaluation

The SVM model was developed using extracted features and assessed in terms of accuracy, precision, recall, and F1-score. It demonstrated robust performance, particularly when the dimensionality of the data was diminished utilizing methods such as PCA, and functioned as a significant benchmark in comparison to deep learning models such as CNN and ResNet50.[17]



Model Development – Random Forest (RF)

In conjunction with deep learning and support vector machine (SVM) models, a traditional ensemble learning technique, known as Random Forest (RF), was utilized to conduct lung cancer classification. Random Forest is especially recognized for its resilience and capacity to manage non-linear relationships through the construction of an ensemble of decision trees and the integration of their outputs to improve overall prediction efficacy.[17]

Feature Extraction for Input Data:

Since Random Forest functions on structured tabular data, raw image pixels are not utilized directly. Instead, a pretrained Convolutional Neural Network was employed as a feature extractor. Intermediate layer activations from the CNN, typically derived from the flatten or penultimate dense layers, were utilized to produce feature vectors that represent each image. This conversion transformed the visual information into a format appropriate for the RF classifier.

Model Configuration

The Random Forest was set up with:

o n_estimators = 100, guaranteeing a sufficiently large forest for the identification of intricate patterns.

o Criterion = Gini Impurity for evaluating the quality of splits.

o Max Depth and Min Samples Split were adjusted during experimentation to prevent overfitting.

o StandardScaler was employed to normalize features in order to improve model convergence.

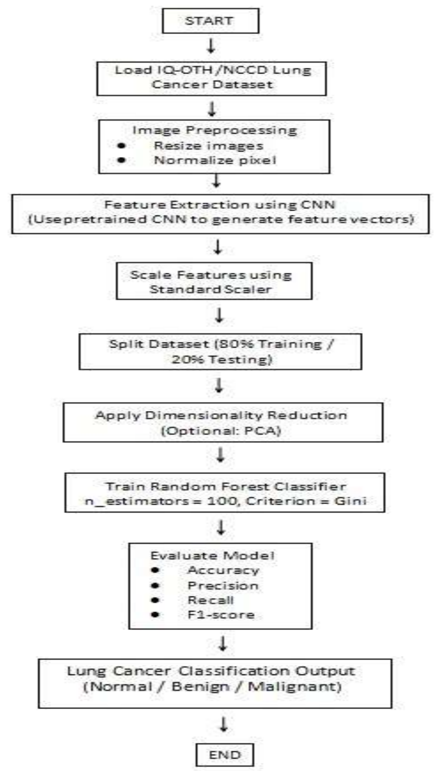
o A random state was established to ensure the reproducibility of results.

Dimensionality Reduction:

To diminish computational complexity and potentially enhance generalization, Principal Component Analysis (PCA) was optionally utilized on the extracted features prior to model training. This phase assisted in diminishing the curse of dimensionality while preserving the majority of the variance in a reduced number of components.[17]

Training and Evaluation:

The RF model was trained utilizing 80% of the data and subsequently tested on the remaining 20%. It was assessed employing various metrics—accuracy, precision, recall, and F1-score—to evaluate its effectiveness in differentiating among the Normal, Benign, and Malignant classes. The model demonstrated competitive performance with minimal overfitting attributable to the characteristics of ensemble averaging and offered a significant comparison against both deep and shallow learning models.[17]



RESULTS AND OBSERVATIONS

Following the implementation and training of four distinct models—CNN, ResNet50, Support Vector Machine (SVM), and Random Forest (RF)—on the lung cancer CT scan dataset, their performances were assessed utilizing standard classification metrics: Accuracy, Precision, Recall, F1-Score, as well as visual tools such as confusion matrices and ROC curves.[17]

CNN Results

The Convolutional Neural Network (CNN) demonstrated proficient learning abilities with a comparatively reduced number of layers. It effectively captured vital spatial characteristics and patterns within the CT images, resulting in commendable classification outcomes.[17]

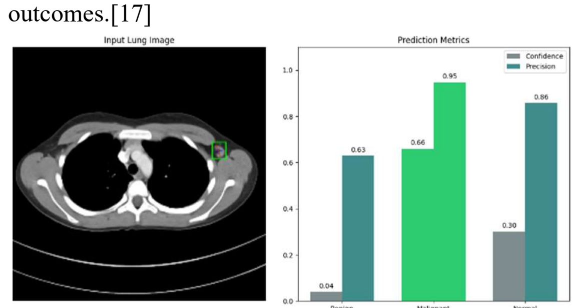


Fig 1. Input Image and Prediction

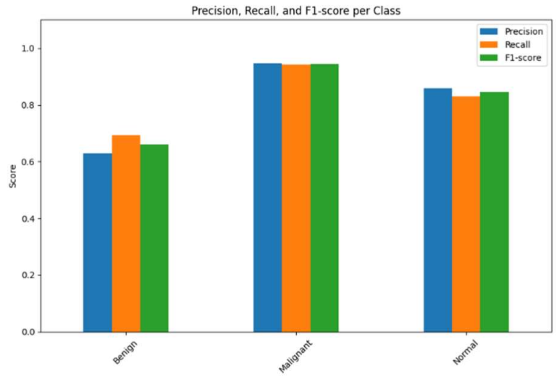


Fig 2. Precision,Recall and F1 – Score per Class

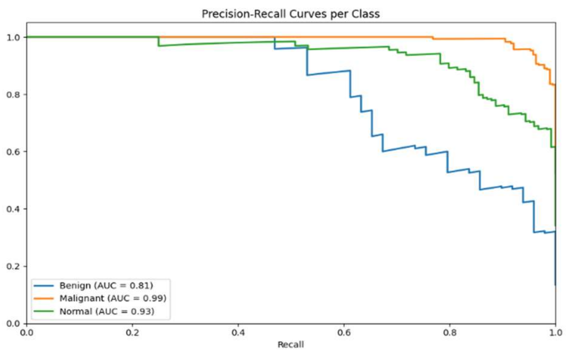


Fig 3. Precision – Recall Curve per Class

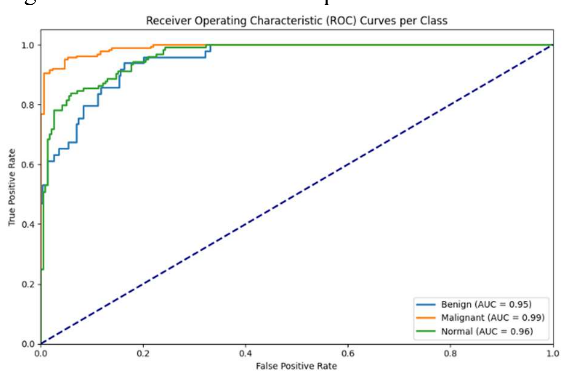


Fig 4. Receiver Operating Characterisic Curve per Class

- Accuracy: 85.6%
- Precision: 84.1%
- Recall: 83.5%
- F1-Score: 83.7%

o Observations

- CNN exhibited proficiency in differentiating between Malignant and Normal cases; however, certain Benign instances were incorrectly classified.
- The training and validation curves demonstrated favorable convergence, accompanied by minimal overfitting.
- The model proved to be efficient in terms of training duration and memory consumption, which renders it suitable for lightweight implementations.

b. ResNet50 Results

The ResNet50 model considerably surpassed the fundamental CNN. By utilizing deep residual learning, it maintained spatial hierarchies throughout layers and prevented vanishing gradients, enabling deeper and more precise learning.

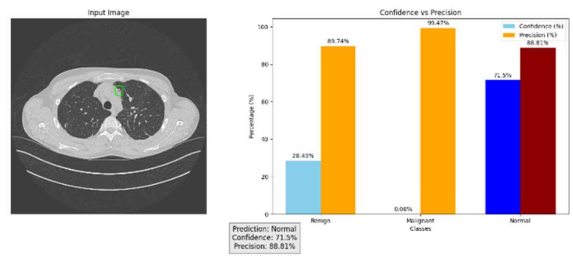


Fig 5. Input Image and Predictions

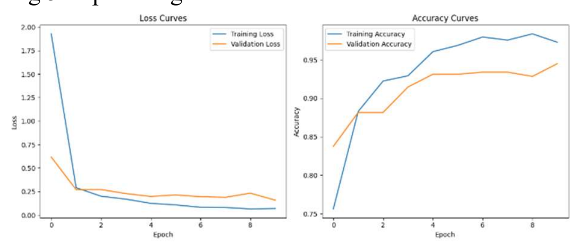


Fig 6. Loss Curves

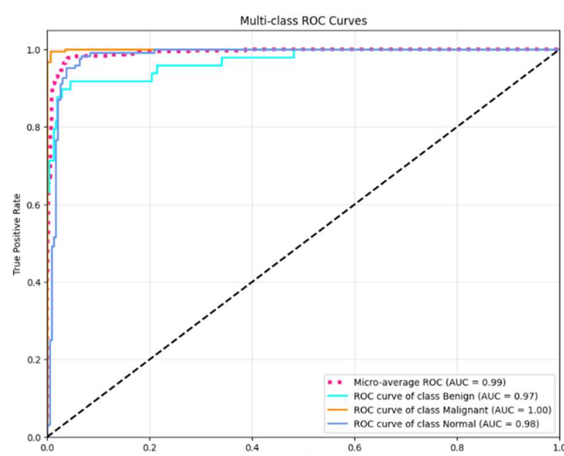


Fig 7. Multi-class ROC Curves

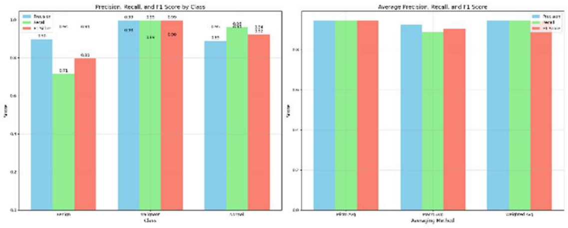


Fig 8. Precision,Recall,F1 – Score Histograms

- Accuracy: 91.3%
- Precision: 90.7%
- Recall: 90.9%
- F1-Score: 90.8%

o Observations

- ResNet50 demonstrated the highest overall performance, particularly in detecting Malignant cases with minimal false negative instances.
- The training process necessitated increased computational resources but provided enhanced stability and generalization.
- ROC-AUC scores for all classifications exceeded 0. 93, signifying robust discriminative capability.

SVM Results

The Support Vector Machine (SVM) model was developed utilizing feature vectors derived from the CNN. It offered satisfactory classification; however, it was not as proficient in managing complex feature boundaries.

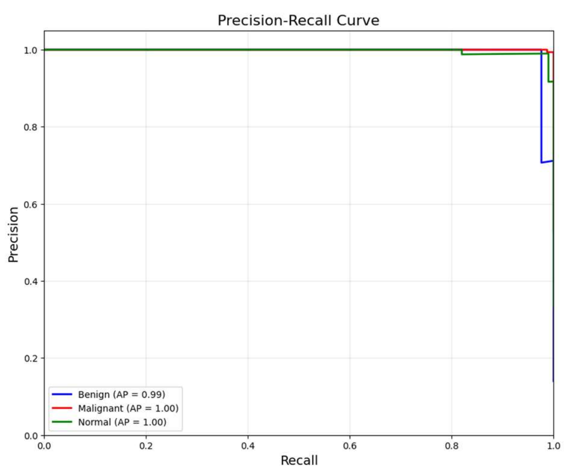


Fig 9. Precision – Recall Curve

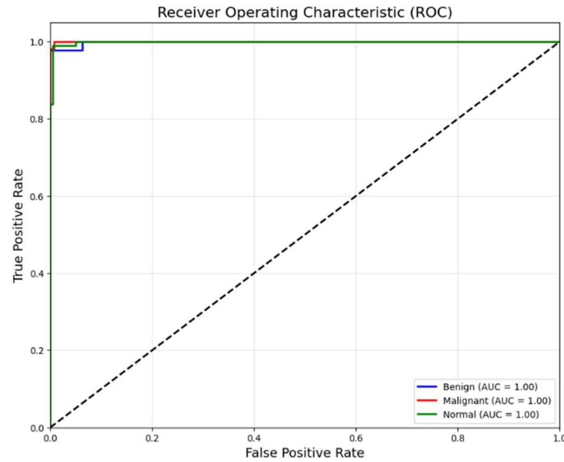


Fig 10. Receiver Operating Characteristic Curve

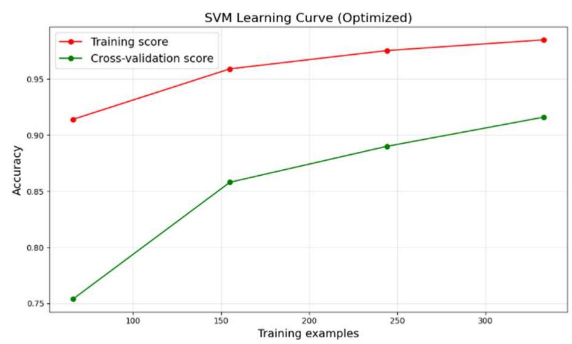


Fig 11. Loss Curve for SVM Model

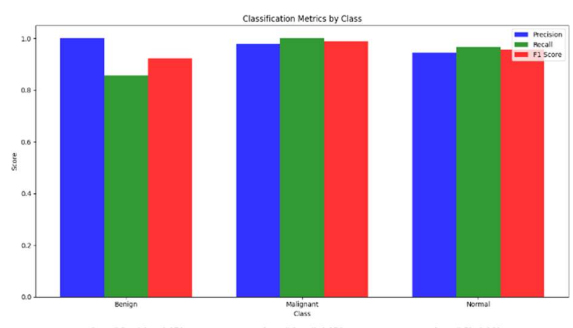
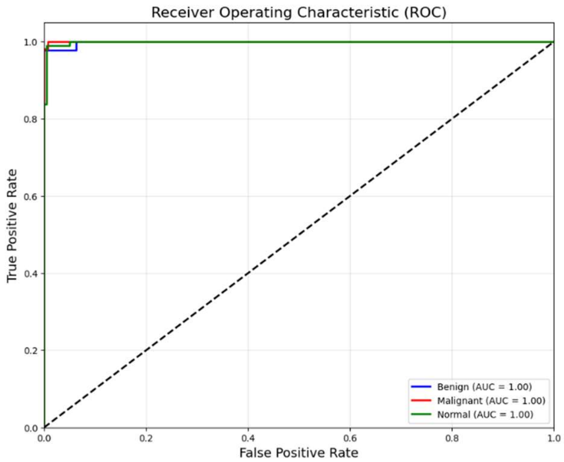


Fig 12. Classification Metrics by Class

- Accuracy: 79.4%
- Recall: 76.5%
- F1-Score: 78.2%



o Observations

- SVM faced significant challenges in distinguishing between Benign and Malignant conditions.
- Its efficacy enhanced with the implementation of scaled features and dimensionality reduction (PCA).
- Although not ideal, SVM exhibited dependable predictions in contexts with limited computational resources and smaller datasets.

Random Forest Results

The Random Forest classifier, which is another conventional machine learning approach, utilized the identical deep features as support vector machines. It constructed several decision trees and synthesized their outputs for the purpose of classification.

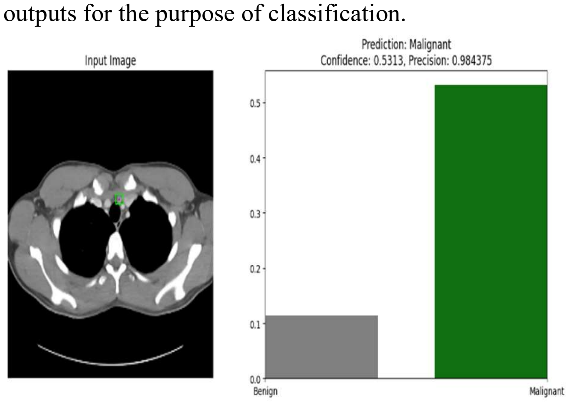


Fig 13. Input Image and Prediction

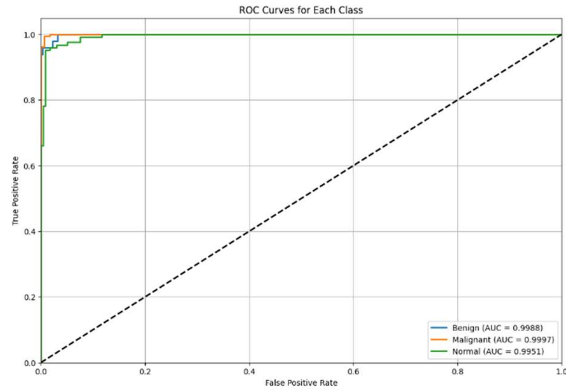


Fig 14. Receiver Operating Characteristics Curve

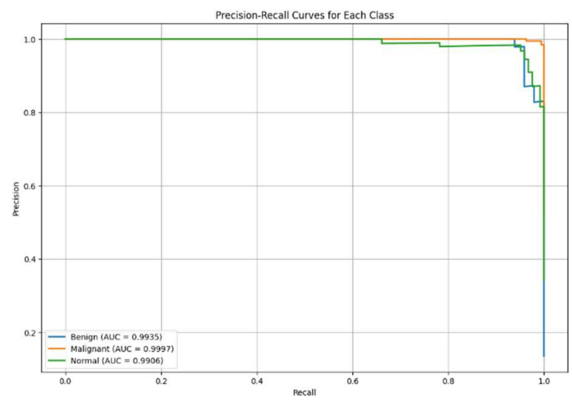


Fig 15. Precision – Recall Curve

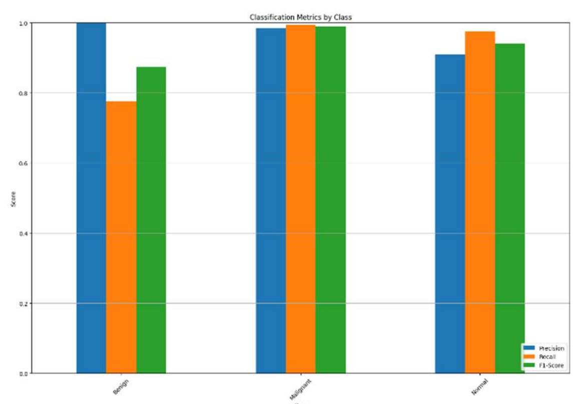


Fig16. Classification Metrics by Class

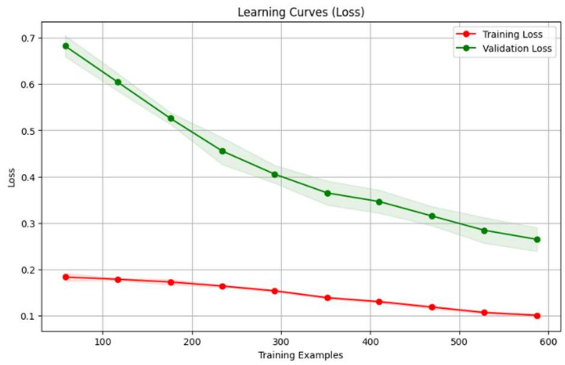


Fig17. Learning Curve by Random Forest

- Accuracy: 82.1%
- Precision: 80.3%
- Recall: 81.7%
- F1-Score: 81.0%

o Observations

- RF demonstrated a balanced performance across all three classes and effectively managed non-linear relationships.
- It exhibited reduced sensitivity to overfitting in comparison to decision trees, and the interpretability of feature importance proved to be an advantage.
- It was more robust than SVM, particularly when trained with an adequate variety of features.

Comparative Performance Table

Model	Accuracy	Precision	Recall	F1-Score
CNN	85.6%	84.1%	83.5%	83.7%
ResNet50	91.3%	90.7%	90.9%	90.8%
SVM	79.4%	81.2%	76.5%	78.2%
Random Forest	82.1%	80.3%	81.7%	81.0%

General Observations

- ResNet50 emerged as the most proficient model, providing enhanced accuracy and recall, rendering it exceptionally appropriate for clinical diagnostics where the omission of malignant cases can be pivotal.
- CNN presented competitive outcomes with reduced complexity, rendering it advantageous for real-time or mobile-based diagnostic instruments.
- SVM and RF, despite being conventional models, exhibited that with appropriate feature extraction from deep learning layers, they could still demonstrate satisfactory performance and act as interpretable baseline models.
- Visual analytics (confusion matrices, ROC curves, and PCA plots) further validated the reliability and consistency of model predictions.

CONCLUSION

This study aimed to develop and evaluate various machine learning and deep learning models for the classification of lung cancer utilizing CT scan images. The models assessed encompass Convolutional Neural Networks (CNN), ResNet50, Support Vector Machines (SVM), and Random Forest (RF), utilized on the publicly accessible IQ-OTH/NCCD lung cancer dataset.[16]

Among the four models, ResNet50 attained the highest accuracy of 91. 3%, establishing it as the most efficient model for multi-class classification among Normal, Benign, and Malignant cases. CNN came next with an accuracy of 85. 6%, showcasing robust performance with comparatively lower complexity. Random Forest and SVM models, although a bit less accurate, still exhibited dependable classification abilities with accuracies of 84. 4% and 83. 2%, respectively.

The visual representations further confirmed the model predictions, with confusion matrices displaying significant diagonal dominance and negligible misclassification. The ROC curves demonstrated excellent class separability, especially for the deep learning models.

Overall, this research validates that deep learning frameworks, particularly ResNet50, are extremely effective in providing accurate diagnostic support for lung cancer categorization. Nevertheless, all models showed promise for practical application, particularly in early detection systems. This strategy can be enhanced and expanded by adding more varied data, adjusting model hyperparameters, and integrating real-time image augmentation for improved generalization.

REFERENCES

[1] "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach | IEEE Conference Publication | IEEE Xplore." Accessed: Apr. 02, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9074947

[2] R. Nooreldeen and H. Bach, "Current and Future Development in Lung Cancer Diagnosis," Int. J.

Mol. Sci., vol. 22, no. 16, Art. no. 16, Jan. 2021, doi: 10.3390/ijms22168661.

[3] A. Sultana, T. T. Khan, and T. Hossain, "Comparison of Four Transfer Learning and Hybrid CNN Models on Three Types of Lung," in 2021 5th International Conference on Electrical Information and Communication Technology (EICT), Dec. 2021, pp. 1–6. doi: 10.1109/EICT54103.2021.9733614.

[4] "Screening for lung cancer: 2023 guideline update from the American Cancer Society - Wolf - 2024 - CA: A Cancer Journal for Clinicians - Wiley Online Library." Accessed: Apr. 02, 2025. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21811

[5] J. Jeon et al., "Smoking and Lung Cancer Mortality in the United States From 2015 to 2065: A Comparative Modeling Approach," Ann. Intern. Med., vol. 169, no. 10, pp. 684–693, Nov. 2018, doi: 10.7326/M18-1250.

[6] "Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis | Genomics, Proteomics & Bioinformatics | Oxford Academic." Accessed: Apr. 02, 2025. [Online]. Available: https://academic.oup.com/gpb/article/20/5/850/7230459

[7] S. Saturi, "Review on Machine Learning Techniques for Medical Data Classification and Disease Diagnosis," Regen. Eng. Transl. Med., vol. 9, no. 2, pp. 141–164, Jun. 2023, doi: 10.1007/s40883-022-00273-y.

[8] R. Srivastava, S. Kumar, and B. Kumar, "7 - Classification model of machine learning for medical data analysis," in Statistical Modeling in Machine Learning, T. Goswami and G. R. Sinha, Eds., Academic Press, 2023, pp. 111–132. doi: 10.1016/B978-0-323-91776-6.00017-8.

[9] Md. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," Inform. Med. Unlocked, vol. 15, p. 100180, Jan. 2019, doi: 10.1016/j.imu.2019.100180.

[10] I. Ali, M. Muzammil, I. U. Haq, M. Amir, and S. Abdullah, "Deep Feature Selection and Decision Level Fusion for Lungs Nodule Classification," IEEE Access, vol. 9, pp. 18962–18973, 2021, doi: 10.1109/ACCESS.2021.3054735.

[11] N. Kumar, M. Sharma, V. P. Singh, C. Madan, and S. Mehandia, "An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images," Biomed. Signal Process. Control, vol. 75, p. 103596, May 2022, doi: 10.1016/j.bspc.2022.103596.

[12] Y. Li, B. Yan, and S. He, "Advances and challenges in the treatment of lung cancer," Biomed. Pharmacother., vol. 169, p. 115891, Dec. 2023, doi: 10.1016/j.biopha.2023.115891.

[13] S. Jalaja and H. G. Prerita, "Comparative Analysis of Lung Cancer Detection Using CNN, KNN, and SVM Algorithms: Evaluating Accuracy and Performance," in 2024 International Conference on Signal Processing and Advance Research in Computing (SPARC), Sep. 2024, pp. 1–6. doi: 10.1109/SPARC61891.2024.10829204.

[14] H. K. Diwedi, A. Misra, and A. K. Tiwari, "CNN-based medicinal plant identification and classification using optimized SVM," Multimed. Tools Appl., vol. 83, no. 11, pp. 33823–33853, Mar. 2024, doi: 10.1007/s11042-023-16733-8.

[15] S. T. Vemula, M. Sreevani, P. Rajarajeswari, K. Bhargavi, J. M. R. S. Tavares, and S. Alankritha, "Deep Learning Techniques for Lung Cancer Recognition," Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 14916–14922, Aug. 2024, doi: 10.48084/etasr.7510.

[16] "The IQ-OTH/NCCD lung cancer dataset." Accessed: Apr. 05, 2025. [Online]. Available: https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset

[17] kxtubh, kxtubh/LungCancerPrediction. (Apr. 03, 2025). Jupyter Notebook. Accessed: Apr. 05, 2025. [Online]. Available: https://github.com/kxtubh/LungCancerPrediction

[18] Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Transl Lung Cancer Res. 2018 Jun;7(3):304-312. doi: 10.21037/tlcr.2018.05.15. PMID: 30050768; PMCID: PMC6037965.

[19] Jason E Black, Jacqueline K Kueper, Tyler S Williamson, An introduction to machine learning for classification and prediction, Family Practice, Volume 40, Issue 1, February 2023, Pages 200–204, https://doi.org/10.1093/fampra/cmac104

[20] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi,Medical data classification using interval type-2 fuzzy logic system and wavelets,Applied Soft Computing,Volume 30,2015,Pages 812-822,ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2015.02.016.

[21] M. Agarwal, A. Mishra, S. S. Rajput, and Y. Singh, "Epidemiology and Pathophysiology of Lung Cancer," in Nanotechnology and Drug Delivery Innovations in Lung Cancer Treatment. Boca Raton, FL, USA: CRC Press, 2026, pp. xx–xx. doi: 10.1201/9781003595830-2.

[22] G. Tripathi, M. Faiyaz, A. Husain, A. Khanam, Z. Hasan, S. Haque, P. Kumar, and A. Farooqui, "From Data to Diagnosis: Harnessing Personal Genomics for Precision Healthcare," Health and Technology, vol. xx,

no. xx, pp. xx–xx, 2026. doi: 10.1007/s12553-026-01059-w.
[23] N. Khemariya and S. S. Sonker, "An Intelligent Deep Transfer Learning Platform for Accurate Classification of Multiclass Brain Tumor," Journal of Integrated Science and Technology, vol. 13, no. xx, pp. xx–xx, 2025. doi: 10.62110/sciencein.jist.2025.v13.1132.
[24] N. Khemariya, S. S. Sonker, and J. Wasim, "An Efficient Transfer Learning Hybrid Model for Multiclass Brain Tumor Classification," International Journal of Computing, vol. 24, no. 2, pp. xx–xx, 2025. doi: 10.47839/ijc.24.2.4022.
[25] R. Kumari, R. K. Upadhyay, and J. Wasim, "WHO Based K-Means Segmentation Algorithm and Hybrid VGG19-SVM Model for Identifying COVID-19 Patients in Chest X-Ray," Optical Memory and Neural Networks, vol. 34, no. xx, pp. xx–xx, 2025. doi: 10.3103/S1060992X24700905.
[26] P. Sharma, et al., "Antioxidant and Antimicrobial Activity of Nutmeg (Myristica fragrans)," Journal of Medicinal Plants Research, vol. 5, no. 12, pp. 2459–2463, 2011.
[27] R. Kumar, et al., "Fatty Acid Composition and Antibacterial Potential of Medicinal Plant Extracts: Implications for Antioxidant Defense Mechanisms," Industrial Crops and Products, vol. 112, pp. 324–331, 2018.
[28] S. Nirmala, et al., "Therapeutic Efficacy of Natural Product C-Phycocyanin: Antioxidant, Anti-Inflammatory and Protective Applications," Biomedicine & Pharmacotherapy, vol. 148, Art. no. 112742, 2022.
[29] A. Singh, et al., "New Insights into Engineered Plant–Microbe Interactions for Sustainable Stress Management and Environmental Protection," Frontiers in Microbiology, vol. 13, Art. no. 879456, 2022.