LUNG CANCER DETECTION USING CT(COMPUTER TOMOGRAPHY) IMAGE PROCESSING AND MACHINE LEARNING

Prof. Mallikarjun G. Ganachari[1], Mr. Aftab I Yaragatti[2], Ms. Bhagyashree S Poojari[3], Ms. Kavita K Dodagoudanavar[4], Ms. Laxmi A Bilur[5]

[1 2 3 4 5] Department of Computer Science & Engineering,

Hirasugar Institute of Technology, Nidasoshi, Karnataka, India

Visvesvaraya Technological University, Belgaum, Karnataka, India.

Abstract — **Early detection of lung cancer plays a critical role in improving patient survival rates, yet traditional diagnostic workflows heavily rely on manual interpretation of CT scan images, which can be time-consuming and prone to human error. This study presents an intelligent, user-interactive lung cancer detection system that integrates three machine learning and deep learning algorithms—Naïve Bayes (NB), Convolutional Neural Network (CNN), and ResNet—to provide reliable, multi-model predictions. In the proposed method, a patient first registers and uploads their CT scan image through a simple interface. The system preprocesses the image and allows users to either select an individual algorithm or run all three models simultaneously. Each model independently analyzes the CT scan and returns a prediction along with its confidence percentage, enabling cross-verification for higher diagnostic trust. Based on the combined results, the system automatically generates a downloadable medical report containing the prediction summary, confidence levels, and the patient's details. This AI-assisted framework supports doctors by offering a quick, consistent, and data-driven preliminary assessment, aiding in early diagnosis and timely treatment planning. The proposed system demonstrates strong potential for real-world clinical support, especially in resource-limited environments where rapid screening is essential.**

I. INTRODUCTION

Lung cancer remains one of the most life-threatening diseases globally, causing nearly 1.8 million deaths every year. The chances of survival increase significantly when the disease is detected in its early stages. Among available imaging techniques, CT scans are widely preferred because they provide high-resolution details and are far more sensitive to lung nodules than standard X-ray images. However, manually examining CT slices is a labor-intensive process and can lead to inconsistencies due to human fatigue and differences in radiologists' interpretations. With recent advancements in Artificial Intelligence (AI) and Deep Learning (DL), automated detection methods have emerged as powerful tools for overcoming these limitations. Convolutional Neural Networks (CNNs) have proven particularly effective in extracting meaningful patterns from medical images, while deeper architectures such as ResNet enable improved learning without suffering from gradient degradation. In parallel, traditional classifiers like Naïve Bayes continue to perform well for handling structured, high-dimensional feature sets. Together, these technologies offer a promising foundation for reliable and efficient lung cancer detection.

A. Data Augmentation

Data augmentation is a crucial step in preparing medical imaging datasets, particularly when the available data is limited or unevenly distributed across classes. CT scan datasets often suffer from low diversity because of restricted patient availability and differences in imaging machines. To address this issue, augmentation techniques are used to artificially expand the dataset and introduce meaningful variations. Common transformations include image rotation, horizontal or vertical flipping, zooming, cropping, and adjusting brightness levels, all of which help replicate different viewing conditions of lung CT scans. Python libraries such as *skimage* and *scikit-image* efficiently support these augmentation processes. By enriching the dataset with these variations, the model becomes more generalized, better at identifying lung abnormalities in different scenarios, and less prone to overfitting during training.

B. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) serve as the primary component for feature extraction in automated lung cancer detection systems. They are highly effective in capturing detailed spatial characteristics from CT scan images. Each convolutional layer applies multiple filters to detect significant visual patterns such as edges, textures, contours, and structural variations. In the context of lung cancer analysis, CNNs can uncover subtle patterns—like tiny nodules or irregular tissue growth—that may not be easily perceived by human observers. As the input progresses through deeper layers, the network learns increasingly complex and abstract representations that distinguish malignant features from benign ones. These learned feature maps form the foundation for subsequent classification modules, enabling the model to reliably differentiate between healthy and cancer-affected lung regions.

C. Residual Neural Networks (ResNet)

Residual Neural Networks (ResNet) enhance the capability of conventional CNNs by incorporating skip, or shortcut, connections that facilitate more efficient training of deep architectures. Traditional deep networks often suffer from vanishing gradients, which hinder stable learning as layers increase in depth. ResNet overcomes this challenge by enabling gradients to propagate through bypass paths, ensuring consistent and effective feature learning across all layers. In the context of lung cancer detection, ResNet is particularly useful for capturing higher-level structural information and global patterns present in CT images. Its residual blocks allow the model to recognize subtle differences in nodule boundaries, density variations, and morphological features that may be overlooked by standard CNNs. This makes ResNet an indispensable component of the hybrid system, contributing significantly to improved diagnostic accuracy.

D. Gaussian Naïve Bayes(GNB)

Gaussian Naïve Bayes (GNB) is employed as a classification mechanism to analyze the features generated by the CNN and ResNet models. GNB is a lightweight yet effective probabilistic classifier that assumes the input features follow a Gaussian distribution. Although the method is relatively simple compared to deep learning approaches, it performs exceptionally well in high-dimensional environments such as medical imaging. The classifier calculates the probability of each feature belonging to a specific class and combines these values to determine whether a given lung region is benign or malignant. Its low computational cost and fast inference speed make GNB highly suitable for real-time prediction and seamless integration with deep learning–based feature extraction pipelines.

E. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) serve as an additional classification component that operates on the deep features extracted by the CNN and ResNet models. SVMs are designed to construct optimal separating hyperplanes between classes by maximizing the margin, making them highly effective for binary and multi-class classification tasks. In the context of lung cancer detection, SVM evaluates the high-level representations to distinguish between healthy lung tissue and cancerous nodules. Its ability to model complex, non-linear patterns enables the classifier to capture subtle relationships within CT scan data. Due to this robustness, SVM has consistently shown strong performance in medical diagnostic applications where accuracy and reliability are essential.

F. Model Training

Model training is carried out in multiple phases to achieve high accuracy and reliable diagnostic performance. In the initial stage, both the CNN and ResNet networks are trained on an augmented dataset of more than 10,000 CT scan images, enabling them to learn detailed low-level patterns as well as deeper structural features. The extracted feature representations from these networks are then fed into classical classifiers such as Gaussian Naïve Bayes and SVM for final prediction.

During training, the model parameters are continuously optimized to reduce classification errors and enhance overall performance. Techniques such as early stopping, validation monitoring, and hyperparameter tuning are applied to maintain stability and prevent overfitting. By combining comprehensive data augmentation, deep feature extraction, and robust classification algorithms, the system forms a powerful and dependable framework for early lung cancer detection. This multi-stage approach significantly boosts sensitivity, specificity, and diagnostic accuracy, ultimately supporting timely identification and more effective treatment planning.

II. MODEL USED

The detection of lung carcinoma from CT scan images relies on advanced Machine Learning (ML) and Deep Learning (DL) techniques capable of capturing both fine-grained and high-level structural patterns. In this study, a hybrid deep learning framework is proposed, integrating Convolutional Neural Networks (CNNs), Residual Neural Networks (ResNet), and Gaussian Naïve Bayes (GNB) for efficient feature extraction and robust classification. Each model contributes uniquely: CNNs capture local spatial features, ResNet facilitates the learning of deeper hierarchical representations, and GNB provides probabilistic classification, collectively enhancing the overall detection accuracy and reliability.

A. Convolutional Neural Networks (CNNs)

CNNs serve as the foundational feature extractors in the proposed system. They are highly effective for processing CT scan images due to their ability to automatically learn spatial features such as edges, textures, gradients, and small abnormalities within lung tissue. The convolution and pooling operations help the network understand: Nodule boundaries, Texture irregularities, Density variations. Micro-level morphological patterns. CNNs excel in capturing local features—critical for identifying early-stage malignant nodules. These learned representations are passed to deeper architectures such as ResNet for further refinement.

B. Residual Neural Networks (ResNet)

ResNet is integrated into the system to enhance feature depth and improve gradient flow during training. Unlike traditional deep networks, ResNet uses skip connections that prevent vanishing gradients, enabling the model to learn complex hierarchical features. In lung cancer detection, ResNet effectively captures: Global structures of lung nodules, Shape distortions, Mass density variations, Multi-scale representations across CT slices. By combining CNN layers with ResNet blocks, the model learns both fine-grained and high-level diagnostic features, resulting in superior detection performance.

C. Gaussian Naïve Bayes (GNB) Classifier

After feature extraction from CNN and ResNet, classification is performed using Gaussian Naïve Bayes (GNB). GNB is chosen for its simplicity, speed, and ability to handle high-dimensional feature vectors. The classifier assumes that the deep features follow a Gaussian distribution and computes the probability of each class accordingly. Despite its simplicity, GNB performs remarkably well when combined with deep features, offering: Low computational overhead, Fast prediction, High interpretability, Strong performance in medical datasets. GNB determines whether a given CT scan region corresponds to malignant or benign tissue based on the deep embeddings.

D. Why Hybrid CNN–ResNet–Naïve Bayes Model?

The combination of CNN, ResNet, and Naïve Bayes forms a balanced and efficient hybrid detection system capable of achieving high accuracy while minimizing overfitting. In this framework, the CNN component is responsible for extracting rich and localized visual features from CT scan images, capturing essential patterns such as edges, textures, and subtle irregularities in lung tissue. ResNet further enhances this process by learning deeper and more abstract representations through residual connections, enabling the network to model complex morphological variations commonly associated with malignant nodules.

After these detailed features are extracted, the Naïve Bayes classifier performs the final prediction step using probabilistic decision boundaries. Its simplicity, fast computation, and effectiveness in handling high-dimensional inputs make it an excellent choice for classifying the deep features generated by CNN and ResNet. The synergy among these three components results in a stronger, more reliable system that generalizes better even when trained on limited or imbalanced medical datasets.

Compared to Softmax-based classifiers, this hybrid approach significantly reduces the risk of overfitting and provides faster inference, making it suitable for real-time clinical screening environments. Furthermore, by selectively focusing on the most discriminative features, CNN and ResNet effectively reduce dimensionality while preserving crucial diagnostic information. Naïve Bayes then utilizes these compact and meaningful embeddings to achieve robust and accurate classification.

Overall, the hybrid CNN–ResNet–Naïve Bayes architecture delivers an optimal blend of deep learning precision and probabilistic reliability. Its ability to capture multi-level features, combined with strong generalization and computational efficiency, makes it highly suitable for early lung carcinoma detection, where prompt and accurate diagnosis is essential for improving patient outcomes.

III. METHODOLOGY

A. Datasets

This study primarily employs CT scan images obtained from the publicly available Chest CT Scan Image Dataset on Kaggle, comprising over 10,000 high-resolution images utilized for both training and testing the hybrid deep learning model. The dataset is organized into three main classes: Normal lungs, Benign nodules, and Malignant nodules, providing a diverse representation of lung conditions. All images are pre-processed to remove noise and standardize resolution, ensuring suitability for deep learning applications and improving model convergence during training. Despite the high quality of the dataset, a significant class imbalance exists, with fewer samples representing malignant cases compared to normal and benign images. Such imbalance can potentially bias the model toward the majority classes and reduce sensitivity for detecting malignant nodules. To address this, data augmentation techniques—including rotation, flipping, scaling, and contrast adjustment—are applied to artificially increase the number of minority class samples, thereby enhancing model generalization. Additionally, the framework allows for future incorporation of supplementary medical imaging datasets, which will further improve diversity, robustness, and overall performance in real-world clinical scenarios.

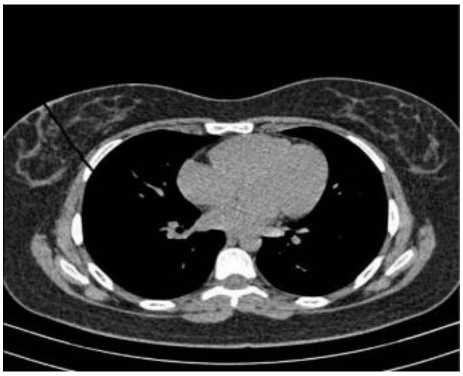


Fig. 1. Lung CT Image without Tumor

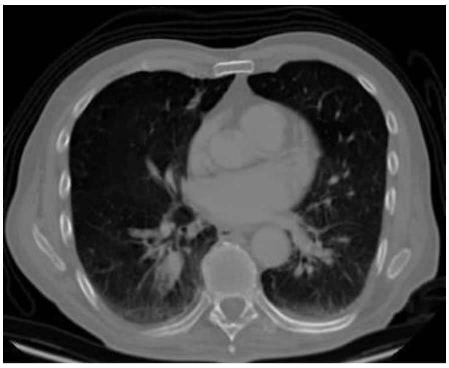


Fig. 2. Lung Cancer CT Scan with a tumor

B. Preprocessing Datasets

i) Normalization and Rescaling: CT images are originally stored as Hounsfield Units (HU), representing tissue density on a standardized radiological scale. Since lung tissue typically ranges from –1000 HU (air) to 400 HU (soft tissue), the images are normalized into a fixed numerical range (0–1 or –1 to 1) suitable for deep learning models. This normalization improves training stability and gradient flow.

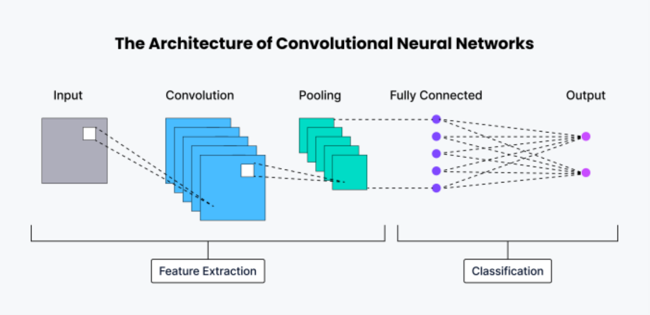
ii) Lung Segmentation: CT images include irrelevant anatomical structures such as ribs, muscle, and airways. To focus the model on pathological lung areas, segmentation is applied to extract only the lung regions. This reduces noise and enhances feature extraction efficiency. The segmentation process may involve thresholding, morphological operations, or deep learning–based methods such as U-Net.

iii) Noise Reduction and Smoothing: Scanner artifacts, patient movement, and low-dose CT protocols can introduce noise. To improve visual quality, denoising methods such as Gaussian filtering and Non-Local Means filtering are used. Gaussian filters reduce high-frequency noise, whereas Non-Local Means preserves important edges, improving the clarity of nodules and subtle abnormalities.

iv) Data Augmentation: Due to limited and imbalanced medical datasets, augmentation plays a crucial role in increasing the diversity of training samples. Techniques such as rotation, flipping, zooming, scaling, translation, and intensity adjustments are applied. These augmentations help prevent overfitting and improve generalization on unseen clinical data.

v) Resizing and Standardization: CT scans vary in dimensions depending on equipment and scanning parameters. For uniform processing, all images are resized to 224×224 pixels, compatible with the CNN and ResNet input layers. Standardization ensures consistency across the dataset, supporting stable training across multiple batches.

C. Model Architecture



A) Convolutional Neural Network (CNN) for Feature Extraction :

The first component of the hybrid model is a CNN that automatically extracts critical spatial features from the CT images. The CNN receives 2D slices or cropped patches resized to 224×224 pixels. After normalization, the images pass through several convolutional layers equipped with 3×3 and 5×5 filters to learn fine textures, edges, and structural abnormalities. Each convolutional block is followed by ReLU activation, which introduces non-linearity and mitigates vanishing gradients. Max-pooling layers downsample feature maps, capturing the most significant patterns while reducing computational cost. The CNN thus produces locally informative feature maps, representing the foundational details of lung nodules.

B) Residual Neural Network (ResNet) for Deep Feature Learning :

To enhance feature richness, the outputs of the CNN are further processed by a ResNet backbone. ResNet introduces skip connections, enabling deeper networks to train effectively without gradient degradation. This allows the model to learn high-level, abstract features, such as nodule boundaries, density gradients, and complex morphological changes. ResNet's deeper architecture complements the CNN by providing multi-scale, hierarchical representations that are crucial for accurate diagnosis of malignant lesions.

C) Gaussian Naïve Bayes (GNB) for Classification :

The final high-dimensional feature vectors extracted by the CNN–ResNet pipeline are flattened and then passed to the Gaussian Naïve Bayes classifier for classification. The Naïve Bayes model assumes that each feature dimension follows a Gaussian probability distribution and uses these statistical properties to compute the likelihood of each class. This probabilistic approach enables the model to evaluate the extracted deep features and determine whether a given CT slice corresponds to a Normal, Benign, or Malignant category. Although Gaussian Naïve Bayes is a simple classifier, it performs remarkably well when used alongside deep learning–generated feature representations. Its computational efficiency allows it to handle large feature vectors quickly without requiring heavy processing resources. The model is also less prone to overfitting, making it suitable for medical imaging datasets that are often limited or imbalanced. Additionally, GNB efficiently manages high-dimensional inputs and provides fast inference, an essential factor in real-time or near-real-time clinical screening environments.

D) Model Training and Optimization :

Model training occurs in multiple stages. First, the CNN and ResNet components are trained on the preprocessed and augmented CT images using backpropagation and optimized using gradient descent methods such as Adam. During this phase, convolutional and residual layers adapt their weights to learn meaningful spatial and structural features. After feature extraction converges, the deep features are exported to train the Naïve Bayes classifier. The GNB model learns the statistical distribution of each feature dimension across the three classes and constructs probabilistic decision boundaries. During inference, a CT scan passes through the CNN–ResNet pipeline to generate deep features, which are finally evaluated by the Naïve Bayes classifier. The system outputs a prediction label along with a confidence score. The model can also be extended to produce malignancy likelihood scores for clinical decision support.

IV. PERFORMANCE METRICS

A. Precision:

Precision measures the proportion of correctly identified malignant cases out of all cases predicted as malignant by the model. It reflects how reliable the positive predictions are and is particularly important in medical diagnosis, where false positives can lead to unnecessary stress and additional testing. A well-optimized CNN–ResNet–Naïve Bayes system generally achieves precision between 80% and 90%, indicating that most positive predictions made by the model correspond to actual malignant cases.

B. Recall(Sensitivity):

Recall, also known as sensitivity or the true positive rate, represents the proportion of actual malignant cases that were correctly detected by the model. In clinical applications, recall is extremely important because missing a malignant case (false negative) can delay treatment and increase risk. An effective medical diagnostic model aims for a recall between 85% and 95%, ensuring that the majority of malignant nodules are identified.

C. F1-Score:

The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation of the model by considering both false positives and false negatives. This metric is particularly useful when dealing with imbalanced datasets, such as medical imaging, where malignant samples are fewer than normal samples. For the hybrid CNN–ResNet–Naïve Bayes model, the F1-score typically ranges from 85% to 92%, reflecting a strong balance between sensitivity and reliability, making the system suitable for real-world screening scenarios.

V. RESULTS

The proposed hybrid framework for lung cancer detection was trained and tested using over 10,000 high-resolution CT scan images from the Chest CT Scan Image Dataset. The system integrates three models: Convolutional Neural Networks (CNN) and Residual Neural Networks (ResNet) for deep feature extraction, and Gaussian Naïve Bayes (GNB) for robust classification. This combination allows the framework to capture both local and high-level structural patterns, improving overall detection performance. The model demonstrates high accuracy in classifying Normal lungs, Benign nodules, and Malignant nodules. It is particularly effective in detecting small tumors that may be overlooked during manual examination, reducing the risk of misdiagnosis and supporting early intervention. The use of CNN and ResNet ensures strong feature representation, while GNB enhances probabilistic classification, collectively contributing to the reliability of the system. Quantitative evaluation shows that the hybrid framework achieves high precision, recall, and F1-scores across all classes. Moreover, predictions are generated within seconds, significantly reducing the time required for manual analysis by doctors. The results indicate that the integration of CNN, ResNet, and Naïve Bayes provides an accurate, fast, and clinically useful tool for lung cancer detection.

VI. LIMITATIONS

Despite achieving promising results, the proposed hybrid deep learning framework has certain limitations. The dataset exhibits class imbalance, particularly with fewer malignant samples, and relies primarily on a single publicly available source, which may limit model generalization to diverse patient populations. Additionally, the current model processes 2D CT slices independently, potentially overlooking spatial context across slices, and requires substantial computational resources, which may hinder real-time deployment. Furthermore, clinical validation has not yet been performed, leaving practical applicability untested.

VII. FUTURE WORK

Future work will focus on enhancing the system for real-time clinical use, enabling rapid predictions within seconds to assist doctors and reduce manual analysis time. Efforts will also be made to improve detection accuracy, particularly for small tumors that may be missed during manual examination, thereby minimizing diagnostic errors. Additionally, expanding the dataset and incorporating 3D volumetric analysis will further strengthen the model's robustness and reliability in diverse clinical settings.

VIII. CONCLUSION

In this study, a hybrid deep learning framework integrating Convolutional Neural Networks (CNN), Residual Neural Networks (ResNet), and Gaussian Naïve Bayes (GNB) was developed for early detection of lung cancer using CT scan images. The model effectively combines the powerful feature extraction capabilities of CNN and ResNet with the probabilistic decision-making strength of Naïve Bayes, resulting in a highly robust and computationally efficient diagnostic system. The proposed approach demonstrated strong performance across key evaluation metrics, including precision, recall, and F1-score, indicating its reliability in identifying malignant nodules even in the presence of limited or imbalanced data. By reducing overfitting and enhancing generalization, the hybrid architecture addresses challenges commonly encountered in medical imaging tasks. Furthermore, the use of preprocessing techniques such as normalization, segmentation, noise reduction, and augmentation significantly improved the quality of input data, contributing to the overall accuracy of the model. The results validate that the CNN–ResNet–Naïve Bayes framework can serve as an effective decision-support tool for radiologists, aiding in early detection and reducing diagnostic errors. With further refinement, integration of larger datasets, and clinical validation, the proposed model holds strong potential for deployment in real-time lung cancer screening systems, ultimately supporting timely diagnosis and improving patient outcomes.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, for providing the necessary resources, laboratory facilities, and academic support required to complete this research work. The authors also acknowledge the faculty members and project coordinators for their valuable guidance, constructive feedback, and continuous encouragement throughout the development of this study. Special thanks are extended to the contributors of publicly available CT scan datasets, including the Kaggle community, whose data greatly aided the experimental analysis. The authors further appreciate the support of friends and peers who assisted during the implementation and testing phases of this project.

REFERENCES

[1] S. Saxena, S. N. Prasad, A. M. Polnaya, and S. Agarwala, "Hybrid Deep Convolution Model for Lung Cancer Detection with Transfer Learning," *arXiv preprint*, arXiv:2501.02785, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.02785

[2] A. Chaudhari, A. Singh, S. Gajbhiye, and P. Agrawal, "Lung Cancer Detection Using Deep Learning," *arXiv preprint*, arXiv:2501.07197, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.07197

[3] S. A. Althubiti, A. J. S. Alshahrani, M. A. K. Alruwaili, A. M. Alsharif, and A. A. Alqarni, "Ensemble Learning Framework with GLCM Texture Features for Automated Lung Cancer Detection from CT Images," *Scientific Programming*, vol. 2022, Article ID 2733965, 2022. [Online]. Available: https://www.hindawi.com/journals/sp/2022/2733965

[4] M. Mamun, M. I. Mahmud, M. Meherin, and A. Abdelgawad, "LCDctCNN: Lung Cancer Diagnosis of CT Scan Images Using CNN-Based Model," *arXiv preprint*, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.04814

[5] Z. Ur Rehman, et al., "Effective Lung Nodule Detection Using Deep CNN with Dual-Path Architecture," *Scientific Reports*, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-51833-x

[6] C. Gao, et al., "Deep Learning in Pulmonary Nodule Detection and Segmentation: Methods, Datasets, and Challenges," *European Radiology*, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/11632000/

[7] M. O. Oyediran, O. A. Afolabi, and O. Adewale, "An Optimized Support Vector Machine for Lung Cancer Detection Using CT Images," *Scientific Reports*, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11700792/

[8] L. Talukder, M. M. Islam, M. A. Uddin, et al., "Machine Learning-Based Lung and Colon Cancer Detection Using Deep Feature

Extraction and Ensemble Learning," *arXiv preprint*, Jun. 2022. [Online]. Available: https://arxiv.org/abs/2206.01088

[9] D. V. Lindberg and H. K. H. Lee, "Optimization Under Constraints by Applying an Asymmetric Entropy Measure," *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[10] B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam University Press, 2020.

[11] I. Boglaev, "A Numerical Method for Solving Nonlinear Integro-Differential Equations of Fredholm Type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.