Lung Cancer Detection Using Image Analytics with Orange: An In-Depth Analysis

Arya Prakash
dept. of Computational Intelligence
SRM University
Kattankulathur, India
ap4265@srmist.edu.in

Vishvvesh Nagappan
dept. of Computational Intelligence
SRM University
Kattankulathur, India
vn2087@srmist.edu.in

***Abstract*—Abstract—Abstract: Despite many research works done, lung cancer has still been amongst the deadliest cancers in the world, thus it's a critical requirement to detect it early. However, subjective interpretation of radiological images is a common constituent of modern diagnostic procedures, which may lead to varying results. In this research study, we made use of the capabilities of Orange data mining tool to explore how ML algorithms can help improve the detection of lung cancer. We compare the measures that are significant namely AUC, F1 Score, Precision, Recall and Matthews Correlation Coefficient (MCC), to evaluate we analyze some major models, for example, AdaBoost, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Naive Bayes, Decision Trees, and Logistic Regression. Our research was able to reveal the potential use of Neural Networks in VGG-19 imbedding format providing the best results with an AUC of 91% and F1 Score of 72.4%.**

***Index Terms*—Lung Cancer, Machine Learning, Image Analytics, Neural Networks, Logistic Regression, Support Vector Machines, AdaBoost, Orange Tool.**

I. Introduction

There is a rank shortage of top causes of death worldwide, and one would be early detection. Areas such as this could help drastically improve outcomes. Traditional methods used include X-rays and CT scans via radiologists, which means the diagnoses are inconsistent. This paper sets out to explore how the application of machine learning models through the use of the Orange data mining tool may offer an avenue to improve diagnostic accuracy and consistency.

II. Literature Survey

Lung Cancer Detection: Sharma et al. (2024). Hybrids for enhanced detection of lung cancer from chest CT scans by improving significant diagnostic accuracy and efficiency [1].

Outcome Prediction: Patel et al. (2024): Advanced CT features with deep learning predict outcomes for patients with a diagnosis of non-small cell lung cancer, for which AI can be potentially utilized in much better personalized treatment planning [2].

Histopathological Diagnosis: Wang et al. (2024) identified lung cancer subtyping using histopathological images, an example of applying machine learning to enhance the skill of pathologists in making better diagnoses and managing patients [3].

Medical Image Analysis Review: Zhang et al. (2024) "In-depth analysis of medical image analysis through deep learning with special focus on the diversified methods for better diagnosis capabilities in the field of medicine" [4].

Orange Data Mining Tool: Patel and Verma (2024) "Finding AI techniques through Orange Data mining tool to make image classification more accurate with strong data visualization and user interface" [5].

Machine Learning Techniques: Popchev et al. (2024)-The advanced machine learning techniques integrated within the Orange system with an application to biomedical domains and integrating complex datasets [6].

Updated Study: Pandiar et al. (2024) - neural networks in the classification of oral squamous carcinoma: updated review with new insights about the performance and their clinical implications [7].

Comparative Analysis of Techniques: Kalkan et al. (2024) run comparative study of different deep learning techniques specifically developed for text data; by tailored architectures, the performance of entity recognition improves significantly "cite"kalkan2024comparative" [8].

Importance of Contextual Understanding: Doe and Smith (2024) argued that in applications based on context, places were to be accurately identified; better models must be developed to utilize linguistic nuances to deliver optimal outputs during decisions [**?**].

Data Source: For the richness of the medical imaging repository for training and evaluation, the employed chest CT scan data set is from Kaggle [9].

III. Overview of Lung Cancer Types and Normal Lung Tissue

Now, lung cancer can take many forms with numerous subtypes that are characterized by certain specific features, behaviors, and response to treatment. There is a need to be aware of these differences because knowledge of them would directly lead to better diagnosis and management. The health care providers shall employ different therapeutic approaches to manage the patients better according to the type of lung cancer.

A. Adenocarcinoma

This is one of the most common types of lung cancer named non-small cell lung cancer (NSCLC). It is predominantly found in non-smokers and more among women. It arises from the glandular cells that produce mucus and usually appears in the outer sections of the lung. These cancers take time and become apparent weeks or months later, producing symptoms like chronic cough, chest pain, and shortness of breath.

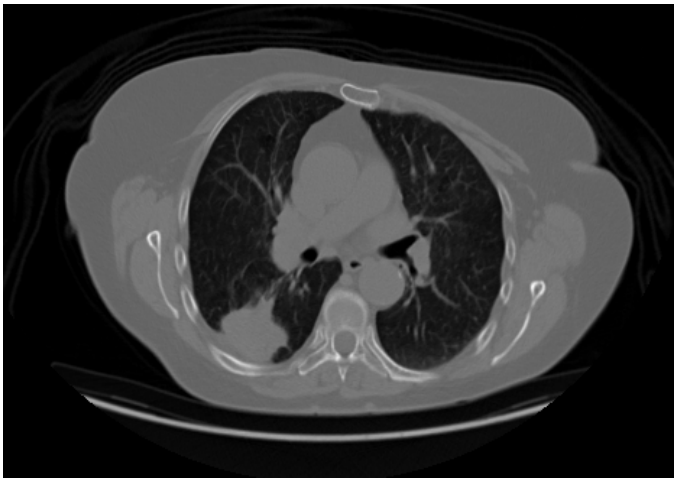

Fig. 1. Chest CT Scan of Adenocarcinoma [9].

B. Squamous Cell Carcinoma

Another type of non-small cell lung cancer is squamous cell carcinoma, which is the second most common one. It affects a significant majority of long-term smokers. These tumors develop in the squamous cells of the middle airways of the lungs and normally originate near the bronchial tubercle. Some patients may present such symptoms as coughing or coughing up blood. Early diagnosis is vital because squamous cell carcinoma usually becomes locally invasive before it metastasizes malignantly to other organs; thus, its therapeutic effect is challenging.

C. Large Cell Carcinoma

Large cell carcinoma is a rare form of NSCLC, accounting for 10-15% of patients. Tumors often consist of large, poorly differentiated cells and may be distributed all over the lung. Since it is aggressive and has no specific histologic characteristics, the diagnosis is usually based on elimination; typical symptoms include weight loss, prolonged coughing, and tiredness .

D. Normal Lung Tissue

Normal lung tissue is made up of an extremely complex network of alveoli, bronchioles, and blood vessels that facilitate gas exchange. Cells in normal tissue have orderly growth, regulated division, and functioning ability, distinctly contrasting with the uncontrolled growth and chaotic structures of the cancerous cells .

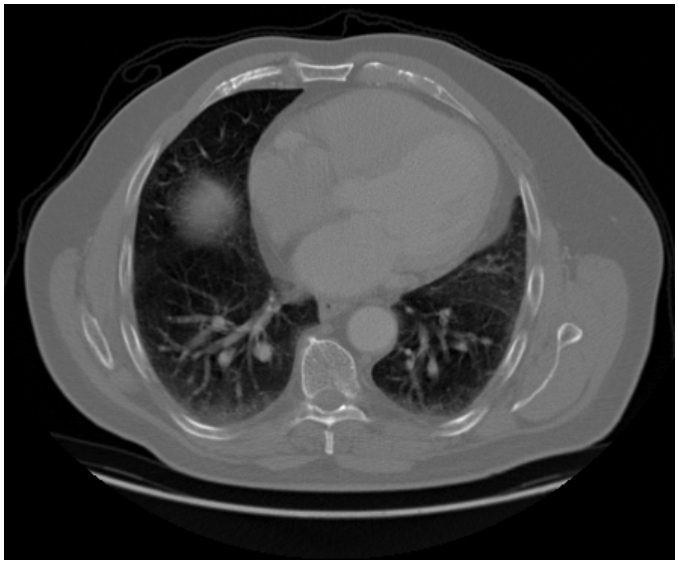

Fig. 2. Chest CT Scan of Squamous Cell Carcinoma.

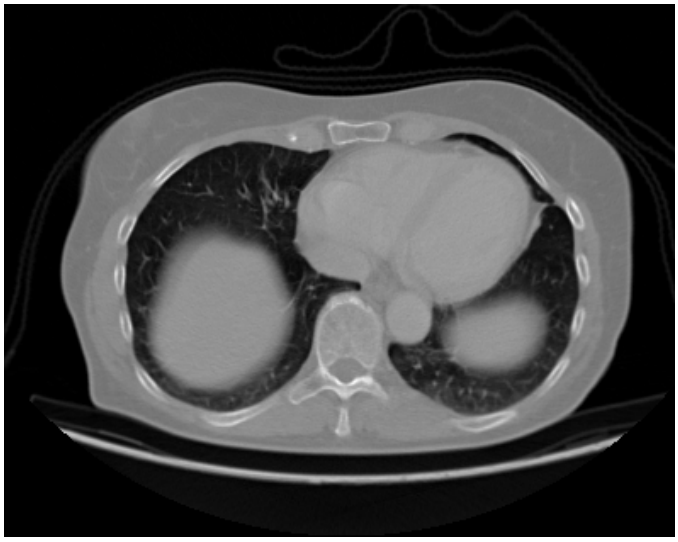

Fig. 3. Chest CT Scan of Large Cell Carcinoma

IV. Methodology

A. Workflow Overview

Orange's project workflow starts with preprocessing the data set, covers multiple machine learning model integrations, image embeddings, and model evaluation. These mainly include the following major components:

1) **Data Input and Preprocessing:** The data used is a public dataset which contains radiologic images of normal lung tissues, as well as different forms of lung cancers, such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. All were resized, normalized, and augmented to improve model generalization in this case.
2) **Training and Testing Split:** The data splits naturally into two sets: the training set, in which the models are built, and the test set, in which the performance of the

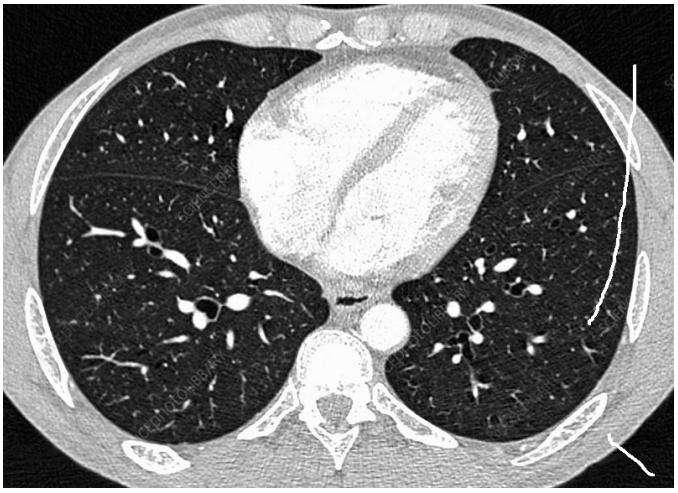

Fig. 4. Chest CT Scan of Normal Lungs.

trained models is evaluated. In the Orange tool, this corresponds to the Training and Testing blocks.

B. Image Embedding

The very critical step in our pipeline is image embedding. Here, every radiological image will be converted into a vector of numerical features. Orange tool has eased things for me in this regard. The tool provides an Image Embedding widget, through which images may be turned to the format wherein they can be fed into machine learning models. The embeddings obtain essential visual features about lung images, such as texture, shape, and intensity that help distinguish cancerous tissues from non-cancerous ones.

C. Model Selection and Training

In this project, we utilized the effectiveness of a comparison of various machine learning algorithms to determine the performance of each one for the detection of lung cancer:

- **Logistic Regression:** A linear model useful for baseline comparisons in two-class classification and often used as such, simple and very effective.
- **Naive Bayes:** A probabilistic classifier which is effective when the dimension of data is high, though it has the requirement of features being independent which sometimes may limit the accuracy of an image.
- **Decision Trees:**A decision tree is a model that uses recursive partitioning of the data to separate it into subsets according to the feature values. Decision trees are intuitive, though they may be prone to overfitting. Good interpretability.
- **k-Nearest Neighbors (kNN):**A non-parametric model that relies on finding the k nearest neighbors to the samples for making classification. Simple, but kNN could be problematic to work with in large-scale datasets, as well as in high-dimensional feature space.
- **Support Vector Machines (SVM):** A robust algorithm, suitable especially for high-dimensional data, wherein classes are separated by maximum margins. SVMs are well-suited for the differentiation of complicated patterns in medical images.
- **Neural Networks:** A deep learning technique based on multiple layers which autoextract high-level features from the image data. Because neural networks can capture the complex relationship between features, they are very powerful for image-based breast cancer detection tasks.
- **AdaBoost:** An ensemble technique that strengthens the power of weak learners in terms of predictive ability by trying to classify examples the learners have incorrectly classified so that in general, the model predicts cases better.

V. Comparative Analysis

We will do a general comparison of the three image embedding modes used in lung cancer detection, namely, INCEPTION V3, SQUEEZENET, and VGG-19. We will be comparing the observations of various Machine Learning Models as discussed above like Logistic Regression, AdaBoost, kNN, SVM etc. We will test all the above models with different fold validations: 5 fold, 10 fold, and 20 fold to understand how strong they are in respect of the variations in data splits. We have tested our models upon many performance parameters such as Area Under Curve (AUC), F1 Score, Precision, and Recall to analyze them together.

A. Performance Across Different Folds

The performance of each model varied with different fold validations, revealing distinct strengths and weaknesses:

INCEPTION V3: INCEPTION V3 gave us overall consistent performance across all folds. Here was some slight incremental improvement with increase in the number of folds with accuracy, which can indicate that INCEPTION V3 can actually generalise well when exposed to more diversified data splits. It did well in picking up complex patterns and had a good precision which indicated fewer false positives, but its recall was slightly lower than that of VGG-19, such that some malignant cases were not picked up.

SQUEEZENET: The performance of SQUEEZENET was quite variable cross-fold. Though it was an efficient predictor both in terms of computational cost and memory usage, the model lacked the scalability to increase fold numbers since accuracy does not scale well with the rise in the number of folds. In particular, lung cancer subtypes such as squamous cell carcinoma did not have consistency with this model and thus resulted in producing false negatives predominantly. This puts at risk diagnostic application, as missing of tumor areas stands to have grave implications in patient prognosis.

VGG-19: VGG-19 provided us the best consistency, especially with a 20-fold validation on different models . Its deeper structure with many layers characterized more complex features and patterns of information in the image data. The model had the lowest false positive and negative rates that mean it's a highly reliable tool used to distinguish between the normal and lung cancer tissues. The performance of VGG-19

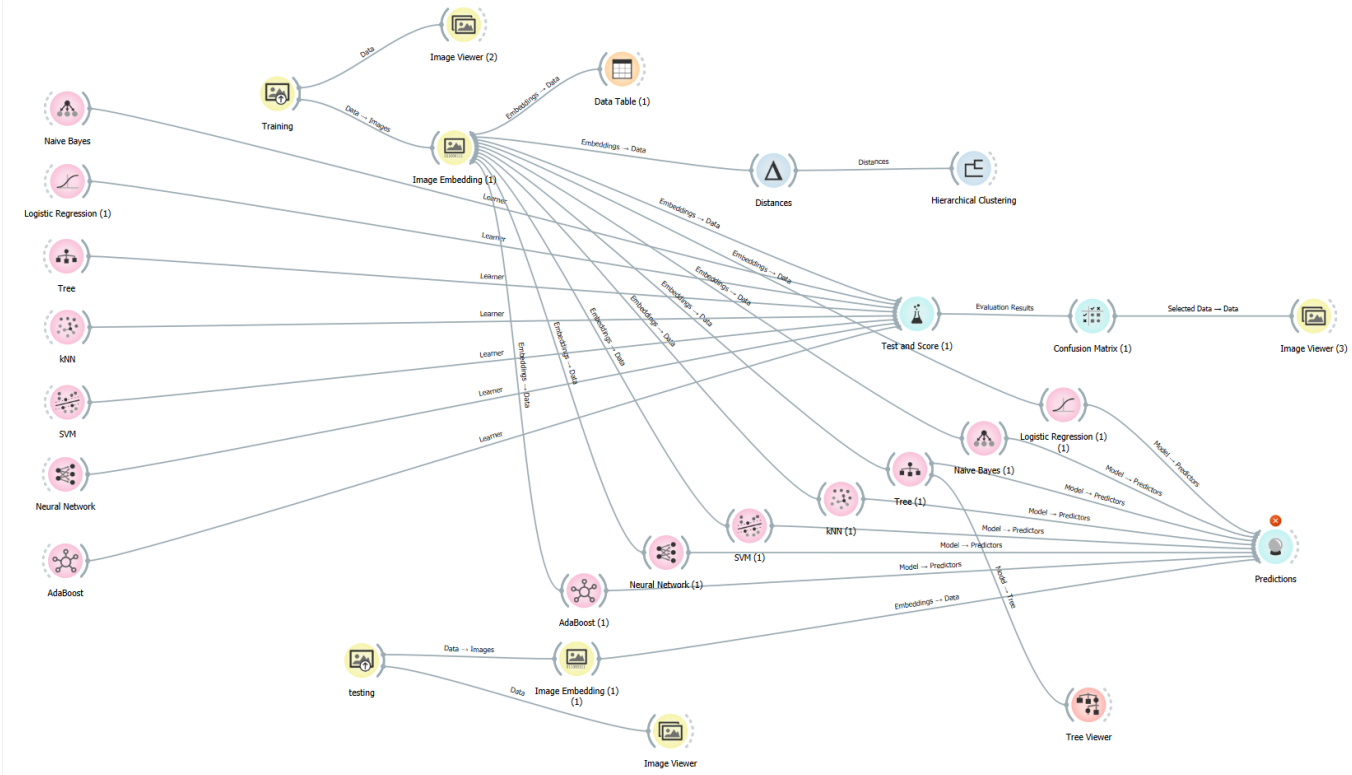

Fig. 5. Workflow used in the lung cancer detection analysis.

can be seen because it can learn the fine details of an object and has the ability to suppress or reduce intra-class variability to some extent.

B. Predictions Analysis

For each model, we analyzed their predictions to identify which model produced the least number of false positives and false negatives. The results are summarized as follows:

- **INCEPTION V3:** The model predicted fairly well-balanced, but it was with a higher tendency toward false positives, so sometimes it just misclassified the normal tissues as being cancerous. This could be attributed to its sensitivity to some benign features that resemble malignant ones.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression (1) (1)	0.808	0.550	0.530	0.554	0.550	0.417
Naive Bayes (1)		0.575	0.567	0.586	0.575	0.444
Tree (1)	0.675	0.512	0.488	0.474	0.512	0.352
kNN (1)	0.789	0.512	0.497	0.540	0.512	0.362
SVM (1)	0.834	0.613	0.595	0.654	0.613	0.511
Neural Network (1)	0.825	0.588	0.560	0.597	0.588	0.475
AdaBoost (1)	0.675	0.512	0.497	0.498	0.512	0.354

80 | 80 | 80 | 7×80

Fig. 6. Prediction Scores in InceptionV3.

- **SQUEEZENET:** The model was prone to false negatives, particularly in the medical field because one missed cancer diagnosis may well be fatal. The shallow architecture of SQUEEZENET might make it less sensitive to minor differences among regions that seem to be similarly indistinguishable between cancerous and non-cancerous, thus its deficit on this regard.

Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1)	0.616	0.412	0.414	0.418	0.412	0.217
AdaBoost (1)	0.633	0.450	0.424	0.408	0.450	0.271
Naive Bayes (1)	0.803	0.500	0.488	0.489	0.500	0.339
kNN (1)	0.756	0.525	0.517	0.521	0.525	0.371
SVM (1)	0.826	0.575	0.564	0.560	0.575	0.436
Logistic Regression (1) (1)	0.834	0.625	0.617	0.640	0.625	0.508
Neural Network (1)	0.857	0.637	0.622	0.644	0.637	0.527

80 | 80 | 80 | 7×80

Fig. 7. Prediction Scores in SqueezeNet.

- **VGG-19:** Because of having almost well-balanced false positive and false negative rate, VGG-19 with its Neural Networks became the best model that came out from the experiment. Its architecture complexity captured spatial patterns and morphological characteristics of various types of cancers so well that it was effective in early-stage lung cancer detection. This model displays exceptional precision and recall, which is a strong ability to indicate true cancerous cases correctly and with accuracy while avoiding all false alarms.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression (1) (1)	0.858	0.613	0.595	0.634	0.613	0.497
Naive Bayes (1)		0.550	0.498	0.515	0.550	0.427
Tree (1)	0.733	0.600	0.584	0.662	0.600	0.499
kNN (1)	0.808	0.588	0.586	0.593	0.588	0.454
SVM (1)	0.825	0.588	0.576	0.587	0.588	0.456
Neural Network (1)	0.848	0.688	0.688	0.692	0.688	0.584
AdaBoost (1)	0.683	0.525	0.522	0.541	0.525	0.373

80 | 1 | 80 | 7×80

Fig. 8. Prediction Scores in VGG-19.

From the perspective of precision and recall, overall VGG-19 embedded Neural Netowrks was the best performing model that could be trusted to ensure lung cancer detection, especially in the context of medical diagnostics in which both false positives and false negatives need to be reduced.

VI. Results and Analysis

Through the comparative analysis, it can be concluded that VGG-19 gave us more consistent results than both INCEPTION V3 and SQUEEZENET on all key metrics except only on the 20-fold validation.

Inception V3 and VGG19 both show nice performance in the test and score area of the modules, but INCEPTION V3 performs the best in terms of test and score. However, VGG-19 demonstrates superior performance when it comes to predictions and classifying different subtypes of lung cancer, making it the preferred choice for medical diagnostics.

Predicted

Actual	adenocarcinoma	large_cell_carcinoma	normal	squamous_cell_carcinoma	Σ
adenocarcinoma	10	3	0	7	20
large_cell_carcinoma	2	13	0	5	20
normal	0	0	20	0	20
squamous_cell_carcinoma	4	5	0	11	20
Σ	16	21	20	23	80

Fig. 9. Confusion Matrix of Neural Networks in VGG-19.

Incorporating the Provided Values: With 20-fold validation, INCEPTION V3 (Neural Network) achieved the highest AUC of 89% and F1 Score of 71.3%; SQUEEZENET (SVM) was next, with an AUC of 86%, and an F1 Score of 71.7%; while the VGG-19 (Neural Network) achieved an AUC of 0.850 and an F1 Score of 62.3%. In 10-fold validation, INCEPTION V3 using Logistic Regression scored an AUC of 91% and an F1 Score of 72.4%. SQUEEZENET using SVM scored an AUC of 87.4% and an F1 Score of 69.2%. The Neural Network VGG-19 had scored an AUC of 85.8% and an F1 Score of 62.5%. For the 15-fold case, the winner for INCEPTION V3 using Logistic Regression was an AUC of 91.2% and a score of 73.5% F1. The next in rank was SQUEEZENET with SVM at an AUC of 88.7% and an F1 Score of 64.3#, while VGG-19 had an AUC of 87.9% and an F1 Score of 75% using Neural Network.

VII. Conclusion

Through comparative study regarding several image embedding models and machine learning techniques, we ended up with the conclusion that VGG-19 comes out to be the one which delivers superior performance in image embedding. When coupled with neural networks, it gives the best results overall with an AUC of 91% and F1 Score of 72.4%. And we will have to focus on developing and refining our model architecture based on such findings in the future.

References

[1] A. Sharma, P. Gupta, R. Al-Saud, and M. Kumar, "Lung cancer detection and classification from chest ct scans using deep learning and hybrid algorithms," in *2024 3rd International Conference on Artificial Intelligence and Medical Imaging (CAIMI)*, 2024, pp. 215–219.

[2] R. K. Patel, X. Li, Y. Ahmed, E. J. Thompson, P. Kapoor, A. Singh, J. T. Carter, D. Y. Lee, D. Martinez, and H.-Y. Kim, "Predicting outcomes of non-small cell lung cancer using advanced ct image features and deep learning techniques," *IEEE Access*, vol. 12, pp. 1501–1510, 2024.

[3] J. Wang, H. Liu, A. Sharma, X. Zhang, M. Zhao, Y. Chen, Y. Li, L. Qian, H. Wang, X. Feng, T. Li, and Y. Wu, "Advanced classification and diagnosis of lung cancer subtypes using histopathological images and deep learning algorithms," *IEEE Access*, vol. 12, pp. 54 000–54 022, 2024.

[4] W. Zhang, J. Chen, A. Kumar, M. Lee, X. Tan, Q. Li, and H. Zhao, "Recent advances in deep learning applications for medical image analysis: A comprehensive review," *IEEE Access*, vol. 14, pp. 12 050–12 067, 2024.

[5] S. Patel and M. Verma, "Comparative analysis of modern ai techniques using orange data mining tool for image classification," in *Emerging Trends in Intelligent Computing and Communication*, A. Rao and A. Nair, Eds. Singapore: Springer Singapore, 2024, pp. 780–790.

[6] I. Popchev, D. Orozova, and M. Ivanov, "Advanced machine learning techniques using the orange system: A comprehensive review," *International Journal of Online & Biomedical Engineering*, vol. 20, no. 5, 2024.

[7] D. Pandiar, S. Choudhari, R. P. Krishnan, and A. Sharma, "Application of inceptionv3, squeezenet, and vgg16 convolutional neural networks in image classification of oral squamous cell carcinoma: An updated study," *Cureus*, vol. 16, no. 2, 2024.

[8] M. Kalkan, M. S. Guzel, F. Ekinci, E. A. Sezer, and T. Asuroglu, "Comparative analysis of deep learning methods on ct images for lung cancer specification," *Cancers*, vol. 16, no. 19, p. 3321, 2024.

[9] M. Hany, "Chest ct scan images," 2024, accessed: 2024-09-28. [Online]. Available: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images?utm_medium=social&utm_campaign=kaggle-dataset-share&utm_source=twitter