Grenze International Journal of Engineering and Technology, June Issue

From Diagnosis to Remission: Understanding Cancer Detection and Staging

Ayan Aggarwal[1], Akshat Bhandari[2], Ayush Singh[3], Ankur Kumar Dubey[4] and Dr Prema R[5]
[1-5]PES University/Computer Science and Engineering, Bengaluru, India
Email: ayanaggar4@gmail.com, {akshatbhandari11, singhayush8638, dubeyankur165}@gmail.com, prema.r@pes.edu

***Abstract*— Lung cancer remains a pervasive threat worldwide, necessitating early detection for effective treatment. Despite the promise of CT scan imaging, challenges persist in accurately identifying cancerous regions and determining cancer stage. To address this, computer-aided diagnosis employing image processing and machine learning, particularly SVM and CNN models, has gained traction. Our study evaluates these models, identifies limitations, and proposes enhancements. While CNN excels in detecting cancerous regions, achieving an impressive 94% accuracy, we augment its capabilities with advanced color-coding techniques to precisely identify tumor regions and classify cancer stages. Our approach aims to significantly improve diagnostic accuracy and clinical decision-making in lung cancer care.**

***Index Terms*— Machine learning, CNN, SVM, CT, SMOTE, Binary classification, Contour analysis, Color-coded representation, Deep learning, Data balancing**

I. INTRODUCTION

Lung cancer poses a significant global health challenge, necessitating timely detection and diagnosis for improved patient outcomes. Traditional methods of lung cancer detection and staging often suffer from limitations in accuracy, efficiency, and accessibility. As such, the development of innovative approaches to enhance diagnostic accuracy and streamline the staging process is crucial.

Our research focuses on addressing these challenges through the development of a web-based Lung Cancer Detection Website. This platform integrates advanced technologies, including machine learning algorithms and image processing techniques, to automate analysis and provide rapid, reliable diagnoses. The objective of this study is to revolutionize lung cancer diagnosis by expediting the detection process and providing valuable insights for optimized treatment plans.

The choice of utilizing a web-based platform for lung cancer detection was motivated by its numerous advantages. Notably, such a system offers scalability, accessibility, and ease of use for healthcare professionals. By leveraging the power of medical imaging and machine learning, our model aims to overcome the limitations of traditional methods and provide actionable information for improved patient management.

The experimental design of our study focuses on seamlessly integrating medical imaging and machine learning algorithms within the Lung Cancer Detection Website. Through comprehensive functionality, including seamless image upload, automated analysis, and insightful report generation, the website aims to enhance the standard of care for lung cancer patients. Our research aims to demonstrate the efficacy of this approach in improving patient outcomes, enhancing healthcare practices, and contributing to advancements in lung cancer care.

Grenze ID: 01.GIJET.10.2.269

II. LITERATURE STUDY

The landscape of lung cancer detection, classification, and staging has witnessed a myriad of approaches, spanning diverse methodologies rooted in both machine learning and advanced imaging techniques. In the quest for heightened diagnostic precision, researchers have explored a multitude of avenues, each contributing to the evolution of the field. [1] introduces a comprehensive and methodical strategy for lung cancer detection and classification through the utilization of advanced CT scan image processing techniques. The approach detailed in the study encompasses a holistic methodology, emphasizing the integration of cutting edge image processing methodologies tailored for CT scans. The methodology encompasses crucial steps, including meticulous image preprocessing, the transformation of preprocessed images into a binary format, feature extraction involving area, perimeter, eccentricity, compactness, and circularity, culminating in effective classification utilizing a Support The landscape of lung cancer detection, classification, and staging has witnessed a myriad of approaches, spanning diverse methodologies rooted in both machine learning and advanced imaging techniques. In the quest for heightened diagnostic precision, researchers have explored a multitude of avenues, each contributing to the evolution of the field. [1] introduces a comprehensive and methodical strategy for lung cancer detection and classification through the utilization of advanced CT scan image processing techniques. The approach detailed in the study encompasses a holistic methodology, emphasizing the integration of cutting edge image processing methodologies tailored for CT scans. The methodology encompasses crucial steps, including meticulous image preprocessing, the transformation of preprocessed images into a binary format, feature extraction involving area, perimeter, eccentricity, compactness, and circularity, culminating in effective classification utilizing a Support Vector Machine (SVM). The paper [2] delivers a comprehensive review on lung nodules, offering profound insights into their definitions, diagnostic evaluations, and management strategies based on existing literature data and guidelines. The categorization of nodules based on size adds a valuable dimension to our understanding of lung nodule characteristics. [3] delves into the realm of lung cancer prediction and stage classification through Convolutional Neural Networks (CNNs). The methodology showcases the utilization of Dicom data and a transfer learning approach, leveraging a pre-trained VGG16 CNN model for enhanced accuracy in classifying different stages of lung cancer. Exploring a more generalized approach, [4] undertakes a comprehensive machine learning perspective for lung cancer prediction. The paper meticulously identifies pertinent features using various selection techniques, subsequently subjecting these features to evaluation through multiple machine learning models, contributing valuable insights into model selection and optimization. [5] concentrates on the pivotal stages of image acquisition, preprocessing, and neural network based classification to achieve efficient and accurate lung cancer detection. This study emphasizes the significance of streamlined preprocessing and effective feature extraction techniques in enhancing the overall accuracy of image-based lung cancer detection. In [6], a multi-stage approach is introduced for lung cancer detection, employing a Multi-Class Support Vector Machine (SVM) based on affected area, Gray Level Co-Occurrence Method (GLCM) for feature extraction, and a watershed based segmentation technique. This sophisticated methodology aims to elevate the precision of lung cancer detection by incorporating multiple stages of analysis. [7] proposes an innovative methodology for the early detection of lung cancer using low-dose CT scans. The method involves a systematic process encompassing preprocessing, fuzzy C-means segmentation, CNN-based feature extraction, and subsequent nodule classification. This systematic approach aligns with the critical goal of early cancer detection. [8] introduces an inventive method for image colorization grounded in isolines and userdefined color constraints. The algorithm minimizes color differences using the Least Squares Method, providing a unique and flexible approach to image colorization.

III. MOTIVATION

Lung cancer remains a pervasive global health concern, with millions of new cases diagnosed annually and a significant portion of cancer-related deaths attributed to this disease. Globally, lung cancer ranks among the leading causes of cancer mortality, with approximately 2.2 million new cases reported each year, resulting in over 1.8 million deaths annually. In India alone, lung cancer accounts for a considerable burden of disease, with an estimated 70,000 new cases diagnosed annually and over 63,000 deaths attributed to this malignancy.

Alarmingly, a substantial proportion of lung cancer cases are diagnosed at advanced stages, limiting treatment options and diminishing survival rates. Moreover, traditional methods employed for lung cancer detection and staging often suffer from inherent limitations, including variability in interpretation and delays in diagnosis.

Our motivation stems from the imperative to address these challenges head-on and empower healthcare professionals with advanced tools for accurate diagnosis and staging of lung cancer. By harnessing the synergistic potential of medical imaging and machine learning, we aim to develop a comprehensive framework that enhances diagnostic accuracy, expedites the staging process, and ultimately improves patient management and treatment outcomes.

Furthermore, the escalating prevalence of lung cancer cases accentuates the urgency of our endeavors. With healthcare systems facing mounting pressure and a growing demand for efficient diagnostic solutions, there is a clear imperative to innovate and devise novel approaches that can meet the evolving needs of lung cancer care.

Through our research, we seek to make substantive contributions to the field of lung cancer detection and staging, with the ultimate goal of translating our findings into tangible improvements in patient outcomes, optimized healthcare practices, and a brighter outlook for individuals affected by this devastating disease.

IV. Problem Domain

At the convergence of medical imaging, machine learning, and oncology lies the focal point of our project: lung cancer detection and staging. Within the expansive realm of healthcare technology, lung cancer diagnosis stands as a complex challenge requiring interdisciplinary solutions. By amalgamating state-of-the-art techniques from medical imaging, notably computed tomography (CT) scans, with advanced machine learning algorithms, our research endeavors to redefine diagnostic accuracy and clinical decision-making in the context of lung cancer care.

V. Problem Definition

The project aims to address the challenge of early detection and accurate staging of lung cancer, which currently poses difficulties due to limitations in accuracy and speed in existing diagnostic methods. Inadequate clarity in medical images and suboptimal tools further exacerbate the issue, while monitoring disease progression remains challenging for healthcare professionals. Our research endeavors to enhance this process by leveraging advanced medical imaging techniques and computer algorithms. By employing specialized imaging modalities and computational tools, we seek to expedite the detection of lung cancer and improve understanding of its severity, thereby enabling more effective treatment strategies.

VI. Statement

This research aims to transform lung cancer diagnosis and staging by integrating advanced medical imaging technologies and machine learning algorithms. By improving the accuracy and efficiency of early detection and staging processes, our objective is to equip healthcare professionals with the necessary tools for timely intervention and optimized treatment strategies, thus addressing challenges related to accuracy, efficiency, and standardization in lung cancer care.

VII. Innovative Content

Our research pioneers ground breaking advancements by fusing leading-edge medical imaging technologies with state-of-the-art machine learning algorithms. Through innovative methodologies, we transcend conventional approaches to lung cancer diagnosis, addressing challenges related to accuracy, efficiency, and standardization. Leveraging advanced imaging modalities and sophisticated computational techniques, we aim to redefine the diagnostic landscape, equipping healthcare professionals with unprecedented insights into early detection and staging of lung cancer.

VIII. Solution Methodologies

In the pursuit of advancing diagnostic accuracy in medical image classification, our research unfolds within the domain of CT image analysis, focusing on the detection of abnormalities. A nuanced and multifaceted exploration is undertaken through the deployment of three distinct deep learning models, labeled Model1, Model2, and Model3, each embodying unique Convolutional Neural Network (CNN) architectures. Additionally, a Support Vector Machine (SVM) is integrated as a baseline model for thorough comparative evaluation. Model1, a sequential CNN architecture, unfolds through convolutional layers followed by activation and maxpooling operations. The model is trained over 10 epochs utilizing categorical cross-entropy loss and the Adam optimizer. Impressively, Model1 demonstrates a high degree of accuracy, particularly excelling in the identification of abnormalities classified under Class 1. A detailed analysis of its training history reveals a progressively decreasing loss, indicating effective convergence. The precision, recall, and F1-score metrics further illustrate its proficiency in achieving high precision for Class 1 abnormalities. Model2, sharing architectural similarities with Model1, undergoes a parallel 10-epoch training regimen. Its performance mirrors that of Model1, underscoring the robustness of the sequential CNN architecture in accurate diagnostic classification. The precision-recall trade-off remains consistent, with an emphasis on precision and high accuracy across various abnormality classes. The validation set results corroborate its effectiveness, particularly in correctly identifying abnormalities belonging to Class 1. Model3 deviates in architecture, presenting a unique configuration within the sequential CNN paradigm. Trained over 5 epochs, it exhibits noteworthy improvements in recall, especially for abnormalities in Class 0. This emphasizes the adaptability of CNN architectures, with Model3 catering to scenarios where achieving higher recall is of paramount importance. Its validation set results highlight its proficiency in accurately detecting abnormalities across diverse classes, particularly excelling in the identification of Class 0 cases. In contrast, the traditional SVM serves as a benchmark for comparative analysis. While maintaining high overall accuracy, the SVM manifests nuanced performance, showcasing a marginal compromise in recall for abnormalities in Class 0. This distinction underscores the importance of tailoring model selection to the specific diagnostic requirements of medical image classification tasks. In systematically evaluating models on a validation set, precision, recall, and F1-score metrics offer a comprehensive perspective. Model1 and Model2, with their shared architecture, emphasize high accuracy and precision, ideal for scenarios demanding precise identification of abnormalities, particularly in Class 1. Model3, with its unique architecture prioritizing recall, emerges as a valuable alternative, excelling in the detection of abnormalities in Class 0. This comparative analysis serves as a guidepost in the development of a robust diagnostic system, facilitating informed decision-making for optimal performance across diverse abnormality classes in medical image classification. The nuanced exploration of different CNN. architectures and the inclusion of a traditional SVM provide valuable insights for tailoring diagnostic models to specific clinical needs, ensuring a comprehensive and accurate analysis of CT images Beyond the technical intricacies of model architectures and comparative analyses, it is imperative to elucidate the real world use case and practical implications of our diagnostic system. The envisioned application of our research lies in augmenting the workflow of radiologists and healthcare professionals, providing them with a powerful tool for swift and accurate abnormality detection in CT images. Our diagnostic system has the potential to significantly expedite the screening process, aiding radiologists in swiftly identifying and prioritizing cases that require immediate attention. This acceleration in diagnostic throughput is particularly critical in healthcare settings where time is of the essence, such as emergency departments or high-volume diagnostic facilities. Moreover, the system's ability to classify abnormalities into specific classes enhances its utility in tailoring treatment plans and interventions, thereby contributing to more personalized patient care.

In addition to the comprehensive diagnostic approach outlined above, our solution methodology incorporates a crucial component for enhancing clinical decision-making: the classification of cancer stages and colorization of tumors. After identifying abnormalities through segmentation and contour analysis, the next step involves determining the stage of the detected tumors based on their characteristics. This process begins with identifying the length of the largest contour, which is indicative of the tumor's size and extent. By analyzing the dimensions of the largest contour, we can categorize the tumors into different stages of cancer progression.

Specifically, the contour radius is compared to predefined thresholds to determine the stage of cancer. If the contour radius is less than 5 cm, the tumor is classified as stage 1. If the radius falls between 3 and 5 cm, the tumor is categorized as stage 2. Similarly, if the radius ranges from 5 to 7 cm, the tumor is designated as stage 3. However, if the radius exceeds 7 cm, indicating that the tumor has spread beyond the chest and potentially to other organs, it is classified as stage 4.

Furthermore, to provide clinicians with a visual representation of tumor staging, we introduce a colorization technique. This colorization process involves assigning specific colors to tumors based on their respective stages. For instance, tumors classified as stage I may be represented by one color, while those categorized as stage II, III, or IV may be assigned different colors accordingly. By color-coding tumors according to their stages, clinicians can quickly identify and differentiate between various stages of cancer progression, facilitating more informed treatment decisions.

This integration of stage classification and tumor colorization enriches our diagnostic system's capabilities, empowering healthcare professionals with a comprehensive understanding of tumor characteristics and cancer progression. Moreover, it enhances the interpretability of diagnostic results, facilitating clearer communication and collaboration among multidisciplinary healthcare teams.

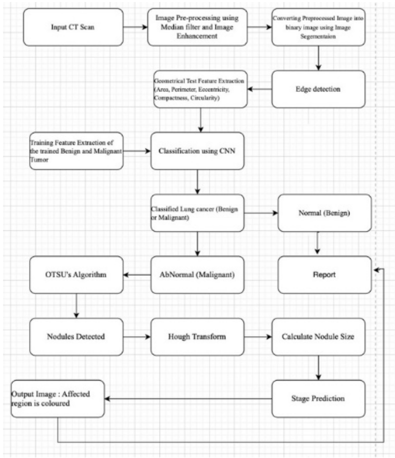

Fig. 1

IX. Results And Sensitivity Analysis

TABLE I: RESULTS

No	Input Image	Input Label	Output Image	Predicted Label	Nodule Size	Cancer Stage
1		Normal		Normal	-	-
2		Benign		Benign	-	-
3		Malignant		Malignant	2.76	1
4		Malignant		Malignant	4.98	2
5		Malignant		Malignant	8.38	3

X. Data Model

Model 1 addressed the challenge of dataset imbalance by applying SMOTE to generate synthetic samples for minority classes, resulting in a balanced dataset. This model utilized a CNN architecture with two convolutional layers, each followed by ReLU activation and max pooling. A Flatten layer was then employed, followed by a Dense layer with 16 neurons. The output layer utilized softmax activation for multi-class classification. On validation, Model 1 achieved an accuracy of 94%.

Model 2 tackled class imbalance differently by computing class weights based on the training data distribution, giving higher importance to minority classes during model training. Despite not employing oversampling or under sampling, this approach led to a highly accurate model, achieving 99% accuracy on the validation set.

For Model 3, the focus shifted to enhancing the model's robustness against variations in input data. This was achieved through data augmentation techniques such as rotations, flips, and cropping, which artificially increased dataset diversity. Although the model's accuracy on the validation data slightly decreased to 93%, Model 3 demonstrated improved resilience to variations in input images, marking a significant advancement in model adaptability.

XI. Comparison Of Results

A. Model Comparison

Based on the evaluation metrics obtained for each model on the test dataset, the summaries of Model 1, Model 2, Model 3, and Model 4 are as follows [1]:

TABLE II. MODEL DETAILS

Sr. no	Model	Accuracy	Precision(macro avg)	Recall(macro avg)	F1 - score(macro avg)
1	CNN with Class Imbalance with SMOTE	0.94	0.89	0.93	0.91
2	CNN with Class Weighted Training	0.99	0.97	0.99	0.98
3	CNN with Data Augmentation	0.94	0.89	0.93	0.91
4	SVM	0.92	0.85	0.91	0.89

B. Different Stages Comparison:

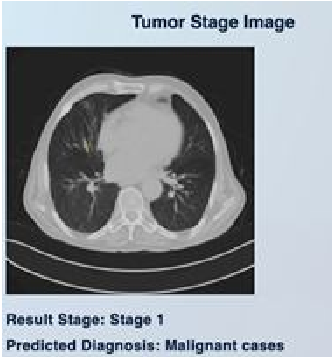

Fig. 3

Fig. 4

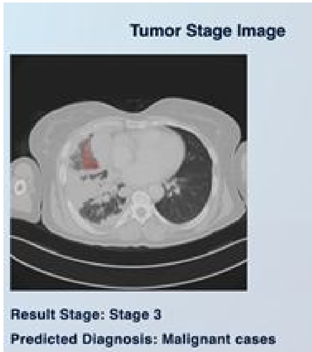

Fig. 5

XII. JUSTIFICATION OF RESULTS

In the evaluation of three distinct models for lung nodule detection, Model 1 employed SMOTE analysis and achieved a commendable accuracy of 0.94. This approach demonstrated a balanced performance across various metrics, showcasing its effectiveness in addressing the challenges associated with imbalanced datasets. Model 2, while achieving an even higher accuracy of 0.99, was excluded from further consideration due to an imbalance in the dataset that resulted in an overemphasis on true negatives. Lastly, Model 3, incorporating data augmentation, achieved an accuracy of 0.94 but was set aside in favor of a more specific and nuanced approach, as data augmentation is a generalization technique for creating synthetic data.
Given these evaluations, we have made the decision to proceed with Model 1. The choice of Model 1 is motivated by its robust performance, striking a balance between accuracy and the nuanced considerations required for reliable lung nodule detection. The utilization of SMOTE analysis in Model 1 aligns with our aim to enhance the model's capacity to detect lung nodules accurately, emphasizing the importance of specificity and sensitivity in medical imaging applications. Moving forward with Model 1, we anticipate gaining valuable insights into predicting lung nodule malignancy and cancer stages with a more focused and tailored approach.
The process of cancer prediction involves the recognition of specific lesions in CT images, considering background evidence of identified risk factors, and assessing the individual's clinical presentation. Moreover, determining the cancer stage is crucial as it provides insights into the size of the tumor and its spread. There are four stages delineating lung cancer progression. For instance, if the size of a nodule (denoted as 'd') is less than 3 cm and hasn't spread to nearby lymph nodes, it corresponds to stage l lung cancer. In cases where the nodule size ranges between 3 cm and 5 cm without spreading to nearby lymph nodes, it is categorized as stage Il lung cancer. Similarly, if the nodule size ranges from 5 cm to 7 cm or beyond, coupled with the spread to nearby lymph nodes, it signifies stage III lung cancer. Finally, when the nodule's radius extends beyond a certain threshold and spreads to multiple sites outside the chest, such as distant lymph nodes or other organs like the liver, bones, or brain, it corresponds to stage V lung cancer. This comprehensive approach integrates lesion detection, size assessment, and lymph node involvement to accurately predict cancer stages, facilitating timely intervention and appropriate treatment decisions.

XIII. CONCLUSION AND FUTURE WORK

In conclusion, this project signifies a remarkable advancement in the domain of medical imaging and cancer diagnosis, showcasing the successful integration of cutting edge technologies, namely image processing and machine learning. Through meticulous development, our system demonstrates proficiency in analyzing CT scan images to detect and categorize lung tumors. Beyond the immediate diagnostic capabilities, our project sets a precedent for the seamless integration of emerging technologies within the medical landscape. The introduction of a color-coded visualization of tumors on lung images not only enhances diagnostic accuracy but also provides an intuitive and informative representation of cancer stages. This contributes to improved communication between healthcare professionals and patients, marking a substantial stride towards more technology-driven healthcare solutions and ultimately fostering enhanced patient outcomes. Moving forward, the future work for this project encompasses several critical aspects. Firstly, the exploration of multi-modality integration, incorporating additional imaging modalities such as MRI and PET scans, is imperative to provide a more comprehensive understanding of tumor characteristics. Furthermore, the investigation into enhanced stage prediction models involves delving into advanced machine learning models and deep learning architectures, aiming for more accurate and nuanced cancer staging. The project's trajectory also involves the initiation of extensive clinical trials and validations to assess system performance across diverse patient populations, ensuring its reliability in real-world scenarios. Establishing a framework for continuous model training, utilizing updated and diverse datasets, is essential to adapt to evolving trends in cancer characteristics. Additionally, the incorporation of genomic data into the diagnostic process stands as a promising avenue, enabling correlation between genetic information and imaging findings for a more personalized approach to cancer diagnosis and treatment. In sum, these future endeavours collectively contribute to the ongoing evolution of technology-driven healthcare solutions, particularly in the context of cancer diagnosis and management.

ACKNOWLDEMENT

The authors wish to convey their sincere appreciation to Dr. Prema R, our esteemed guide, for her invaluable support, guidance, and mentorship throughout the duration of this research endeavor. Furthermore, the authors express their profound gratitude to PES University for providing the requisite resources and fostering an environment conducive to research excellence.

REFERENCEES

[1] Syed Saba Raoof , M A. Jabbar, and Syed Aley Fathima, “Lung Cancer Prediction using Machine Learning: A Comprehensive Approach”, 2020.

[2] V. Deepa, P. Mohamed Fathimal, "Lung cancer prediction and Stage classification in CT Scans Using Convolution Neural Networks - A Deep learning Model", 2022.

[3] Janee Alam, Sabrina Alam, Alamgir Hossain, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier", 2022.

[4] Adam Popowicz, Bogdan Smolk, "Isoline Based Image Colorization", 2014.

[5] Nusrat Nawreen, Umma Hany, Tahmina Islam, "Lung Cancer Detection and Classification using CT Scan Image Processing", Department of Electrical and Electronic Engineering, 2021.

[6] Konstantinos Loverde’s, Andreas Fotiadis, Chrysoula Kontogianni, Marianthi Iliopoulou, "Lung nodules: A comprehensive review on current approach and management".

[7] S.A.D.L.V. Senarathna, S.P.Y.A.A. Piyumal, R. Hirshan, W.G.C.W. Kumara, "Lung Cancer Detection and Prediction of Cancer Stages Using Image Processing", 2021.

[8] B C Kavitha, K. B Naveen, "Image Acquisition and Pre-processing for Detection of Lung Cancer using Neural Network", 2022.

[9] Gagan Thakral, Sapna Gambhir, Nagender Aneja, "Proposed methodology for Early Detection of Lung Cancer with low-dose CT Scan using Machine Learning", 2022.

[10] Mohd Firdaus Abdullah, Muhammad Safwan Mansor, Siti Noraini Sulaiman, Muhammad Khusairi Osman, "A Comparative Study of Image Segmentation Technique applied for Lung Cancer Detection", 2019.