IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Artificial Intelligence In Cancer Detection: A Study On Lung And Colon Histopathology Images

[1]M.Karthikeyan, [2]Dr.K.Dharmarajan

[1]Research Scholar, [2] Professor,
[1] School of computing Sciences,
[1] VISTAS, Chennai, India

Abstract: A global challenge that has become a menace is the example of cancers such as lung and colon cancer (LC AND CC) Colon cancer(CC). The path of this research, therefore, in this very important area, is early detection through the use of AI in histological image analysis. The study introduces a novel hybrid feature set aiming to improve classification accuracy by integrating DenseNet201 with color histogram techniques. The validation of this feature set works with eight major ML algorithms-KNN, SVM, Light GBM, CatBoost, XGBoost, decision trees, random forests, and multinomial naive Bayes. This comprehensive study, therefore, highlights an exotic model that achieved an accuracy of 99.683% on the LC AND CC25000 dataset. Taking the same concept to breast cancer detection using the Break His dataset shows a great accuracy of 94.808%. These results highlight the revolutionary potential of AI in the challenging area of histopathological analysis and therefore stand to become a transformative player in accelerating diagnostic accuracy. A thorough comparative analysis showcases the strengths and weaknesses of current AI practices in medical imaging, outlining a pathway for improvement and future clinical application.

Index Terms **– Densenet201, histopathological images, image processing, lung and colon cancer, machine learning.**

I. INTRODUCTION

Imagine a world in which medical problems such as lung and colon cancer could be diagnosed with a level of precision unheard of before, allowing for treatment to begin earlier for the promotion of longer survival. The scientific realm where this vision is now more attainable than ever is in the AI adaptive learning for improved detection of cancer. The movement to automate the detection system, with hybrid histopathology image meta-analysis, is definitely at the cutting edge of these exploratory advances. This paper will discuss how it has turned the cancer-diagnosing and treating paradigm upside down, its methodologies, and implications for healthcare. For this reason, in 2023 alone, with an expected 1,958,310 new cancer cases facing the United States and 609,820 deaths from cancer, it is perhaps time to surface the reality of the cancer problem this year. Furthermore, this translates to an estimated 5365 new cases and about 1671 cancer-related deaths every day. Thus, even with the auspices of science and technology making fast-paced strides, cancer continues to be a challenge. [1] Among these, lung and colon cancer (LC AND CC) appears as a formidable threat to the progress being made to date, estimated by the American Cancer Society to further with about 238,340 new LC AND CC cases anticipated for this year. [2] Affecting, on average, the ages above 70, it increases the burden for the elderly trying to manage a probably life-threatening disease. Due to being No. 1 killer of people with cancer Globally, effects of LC AND CC are indeed felt hard on healthcare systems, families, and on the economy. With advances in all other fields of medicine, no positive reduction in the incidence of LC AND CC has been registered yet [4], [5]. This is coupled with the urgent demand for novel and effective strategies for early cancer detection and treatment. We intend to tackle this problem using DenseNet201 architecture

and color histograms and combining them with other ML algorithms for improving the accuracy and efficiency of the diagnosis, particularly for the early detection of lung and colon cancer.

The burden caused by LC AND CC is huge, with the latter representing the leading cause of cancer deaths in the U.S.A., particularly LC, contributing approximately 20% of all cancer deaths [6]. This burden exceeds that from colon, breast, and prostate cancers combined, reflecting its tremendous public health importance [2], [7]. Cancer development is influenced by multitudes of behavioral and environmental factors including but not limited to smoking, obesity, or radiation-alcohol consumption(LC), caffeine consumption (CC), racketing health services [8]. The early detection of LC AND CC is also particularly difficult because the diseases often remain asymptomatic or only produce subtle symptoms in early disease stages, which then become responsible for the complete diagnosis delay. By the time that symptoms develop, this cancer is quite often already in the advanced stage compromising any chances for efficacious early intervention [9].

II. The Importance of Histopathology in Cancer Diagnosis

Histopathology involves the microscopic examination of tissue samples to identify disease. It plays a pivotal role in cancer diagnosis, aiding in the determination of tumor type and stage.
Histopathology Basics

Histopathology essentially includes:

- Tissue Sampling: Biopsy or excision of tissues for analysis.
- Slide Preparation: Fixation, embedding in paraffin, and slicing thin tissue sections.
- Staining: Applying various stains to accentuate the cellular structures.
- Microscopic Analysis: Viewing by pathologist in identified abnormalities using stained slides.

The images generated on such slides are usually very informative and can be interpreted manually by trained pathologists. However, the whole procedure can be tedious and can vary in human interpretation

CONTRIBUTIONS

The proposed research is progressive as it uses an atypical amalgamation of features obtained through different methods. This is probably due to the complex and high-dimensional nature of the data, especially in histological images. Precursor abnormal cells frequently show resemblance in their characteristics; hence, hybrid systems combining features from various methods are required. Such a combination will improve models' discriminatory power regarding subtle differences that might indicate the onset of LC AND CC [27]. The following parts summarize the contributions made by the study.

• Feature Extraction Using DenseNet201 and Color Histograms:
This study uses DenseNet201 and color histogram methods for the extraction of deep features from histological images in the LC AND CC25000 dataset. The dataset is concentrated on lung adenocarcinoma, colon adenocarcinoma, squamous cell carcinomas of the lung and colon, and benign lung and colon tissues. Such an approach gives a hybrid feature set highly discriminative among classification models.

• Evaluation of ML Algorithms:
The performance of eight ML algorithms is evaluated-KNN, LGBM, CatBoost, XGBoost, DT, RF, MultinomialNB, and SVM. All these algorithms are evaluated on the parameters of accuracy, specificity, precision, recall, F1-score, and computational efficiency in order to determine the best algorithm for this task.

• Multi-class and Binary Classification Tasks:
The multi-class classification task is conducted to identify lung adenocarcinoma, colon adenocarcinoma, lung squamous cell carcinoma, colon squamous cell carcinoma, and benign classes related to LC AND CC. Binary classification tasks are also introduced between benign versus adenocarcinomas, benign versus carcinomas, and adenocarcinomas versus carcinomas so that more information is provided regarding subtypes of LC AND CC and their distinguishing features

III. METHODOLOGY

This section describes the general process of lung and colon cancer (LC AND CC) classification mentioning the crucial steps that form the base of the overall procedure. The framework has been designed for linear phases, such as selection of dataset, data preprocessing, feature extraction (FE), fusion of features, implementation of machine learning (ML) models, and evaluation of performance. It is indeed an elaborative process where each phase plays an important role in enhancing accuracy as well as efficacy in LC AND CC detection.

3.1LC AND CC25000 DATASET

In this study, we used the LC AND CC25000 dataset, which is a rigorous and detailed dataset of histopathological images for lung and colon cancer (LC AND CC) examination. Obtained from the Kaggle initiative, this dataset was organized by Andrew Borkowski and his associates at James Hospital in Tampa, Florida. The dataset comprises different types of cancers; hence, it includes lung cancer as well as colon cancer and combined lung-colon cancer. From the entire set of 25,000 images in the dataset, we picked 15,000 images from three LC AND CC categories: adenocarcinoma (lung_and_colon_aca), which constitutes most cases of LC AND CC; benign lung and colon tissue (lung_and_colon_bnt); squamous cell carcinoma (lung_and_colon_scc), which is the second most prevalent form. We included 10,000 images from two CCS categories Colon Adenocarcinoma (colon_aca) and Colon benign polyps (colon_bnt). Our focus on these particular types enables a targeted in-depth analysis of the features characteristic of LC AND CC.. Each category includes 5,000 images, ensuring a balanced sample across classes.

Figure 1 presents sample images from the dataset, while Table 2 features a pie chart to illustrate the class distribution. Visualizing the dataset composition is essential for confirming that the model is trained and tested on a balanced and representative sample. Initially, the dataset was generated from 1,250 primary pathology slide images (250 per category) and was augmented through rotations and flipping to create a total of 25,000 images. Each image, originally sized at 1024 × 768 pixels, was resized to a standardized 768 × 768 pixels to ensure consistency in analysis [28].

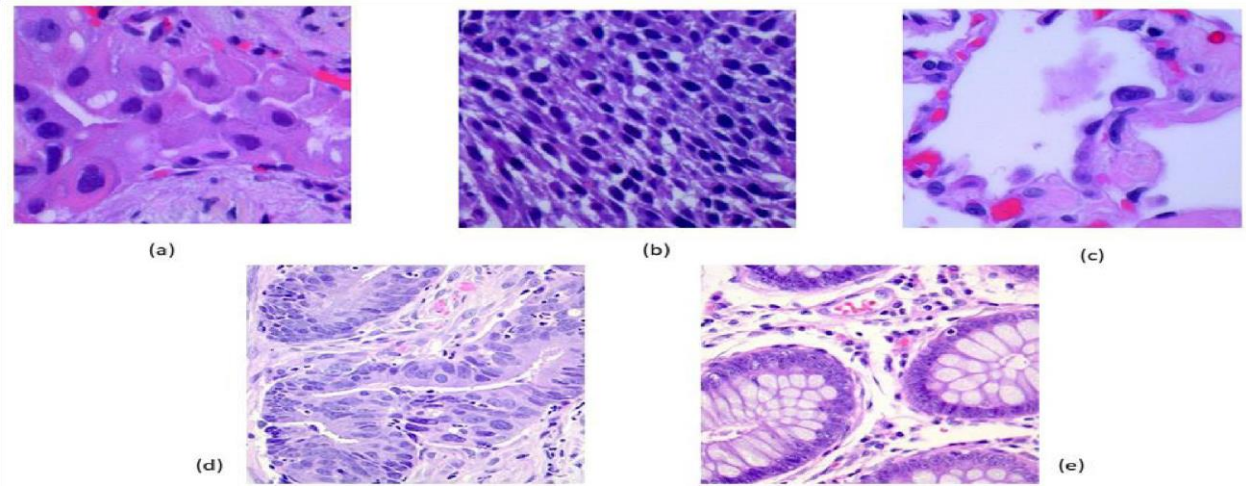


Figure 1 (a) Lung and colon Adenocarcinoma. (b) Lung benig. (c) Lung squamous cell. (d) Colon benig. (e) colon Adenocarcinoma

IV. PREPROCESSING STAGE

The first preprocessing of the dataset is the LC AND CC image dataset. The images are now loaded in RGB mode and resized to 128 × 128 pixels to balance computational efficiency and detail retention needed to be accurate in classifying them. The scaled images and their labels are now converted into numpy arrays to facilitate processing. More so, there is a custom function (ensure_correct_depth) that will ensure the image depth is set to CV_8U so that all images will be scaled in the same way and in the same format to allow the feature extraction to be consistent across the dataset. Here converted to HSV image color space, this is really slicing open the hues characteristics of colorization distribution. Then make the image preprocessing appropriate for model training, the input to LeNet should be processed using the preprocess_input function that ensures normalization, mean RGB subtraction, and standard deviation by the ImageNet dataset. E Step

checkpoints this process for uniformity of the input data distribution with that of the trained model for better performance.

USSION with respect to SAGE Publications has been issued on behalf of and under the authority of the Director of the UK Centre for Research on Globalization under the Hague Conventions. All pages have been marked with their publication dates, which correlate with the actual research dates in the SAGE Publication Limited Scientific Research program. Should new evidence of institutional conditions emerge, SAGE Publications invites re-examination of the current policy. All SAGE Collections have voided. If anyone is interested in applying for enhanced or updated content in any of the SAGE materials, please contact SAGE Sales or visit the relevant link. SAGE has also completed its consolidation under any of the SAGE Collections for any material under such material types

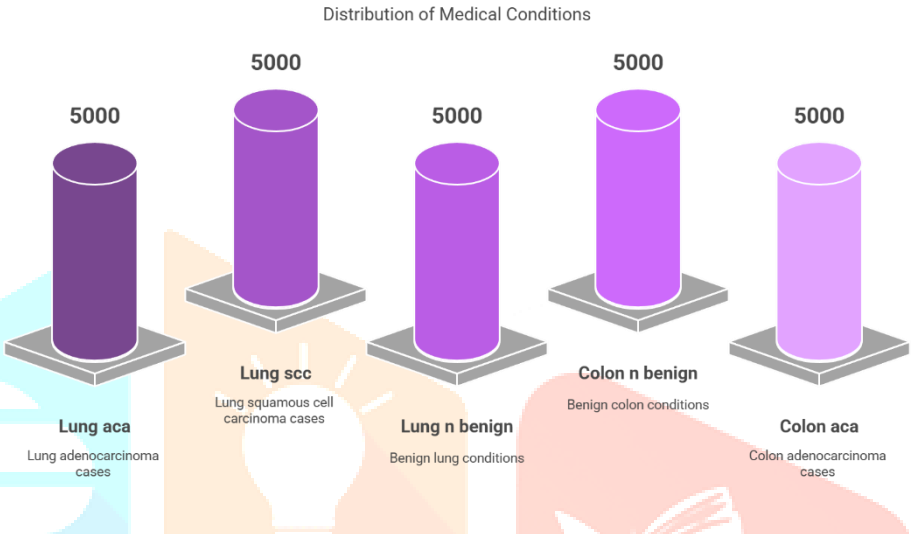


Figure 2 dataset 5 classes

4.1FEATURE EXTRACTION (FE)

The model presented in the developed methodology contains some crucial stages for LC AND CC classification, as can be seen in Fig. 2. Primarily the work relies on feature extraction (FE). This methodology combines image processing and deep learning (DL) techniques to extract relevant information from histopathological images in the LC AND CC25000 dataset.

4.2COLOR HISTOGRAM ANALYSIS

In this phase, the images are first converted to the HSV (Hue, Saturation, Value) color space, widely used for image processing, mainly because it separates the intensity (Value) out of the color information (Hue and Saturation), thus making the method robust against variations in lighting conditions [29]. Then a color histogram is calculated for each image, with special emphasis on the Hue component, which plays a crucial role in distinguishing color-based features in medical images. Understanding color distributions is highly important since these distributions can highlight themselves in subtle ways in different lung and colon conditions. The histograms are then normalized to ensure consistent feature scaling. This is critical in ensuring that features with greater numeric ranges do not overwhelm the classification models. The normalized histogram is then treated as a probability distribution of Hue values, while each image histogram is flattened into a one-dimensional feature vector that is amenable to input into machine-learning models. CONTOUR FE

In the contour extraction step, the images are first converted to grayscale, simplifying the image data and emphasizing structural details [31]. Contours in the grayscale images are then detected, representing regions that may hold medical significance in lung and colon scans. This method highlights the structural patterns that are most likely relevant for medical diagnosis. The contours are approximated to reduce the number of points needed to represent the contours, focusing on the most important horizontal, vertical, and diagonal segments [32]. For each contour, various features are extracted, such as the area and perimeter of the contour, which are then flattened into a feature vector for each image. These feature vectors capture the structural properties of

the lung and colon images, which are crucial for distinguishing between different types of lung and colon cancer in the classification phase.

4.3DENSENET201 FEATURES

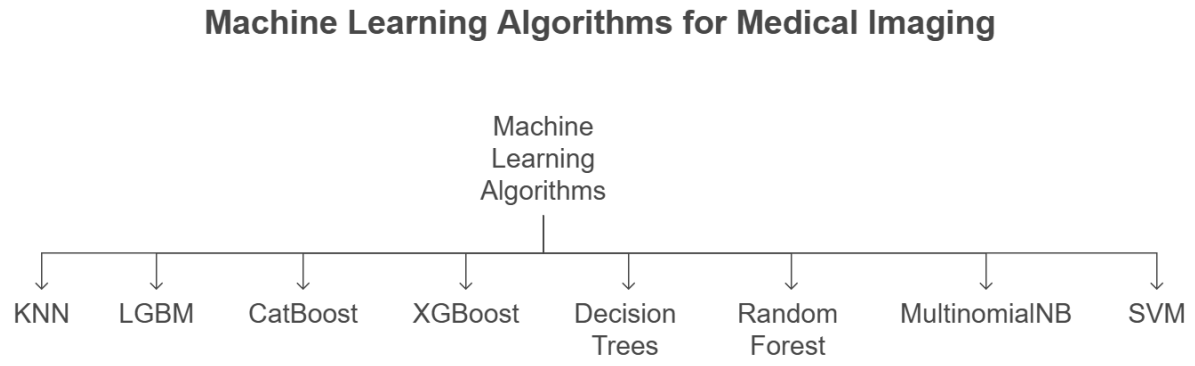
DenseNet201 is a neural network, the architecture of which connects every other layer to any other layer in a feed forward manner in such a way that it optimizes the flow of information between them [33]. Therefore, this DenseNet201 is used with pre-trained weights from the ImageNet database in our methodology to give a benefit for the model to utilize all the knowledge gained from training in such a huge dataset, so it has a great improvement over feature extraction in terms of lung and colon cancer classification. The network is now changed to a feature extractor by removing its top classification layer. The DenseNet201 layers capture features from simple edge descriptions in their initial layers to complicated representations or patterns in the deeper ones in the pipeline as the images go through the network [34]. Global Average Pooling (GAP) is the last layer of pooling behind convolutional layers that reduces the dimensionality of feature maps while preserving important spatial information. Hence, one vector of features is given for every image, which reflects a holistic understanding of the image content in a detailed way, including small to large patterns more relevant for the classification of lung and colon cancers.

4.4FEATURE COMBINATION

The tactics to integrate the extracted attributes are manifold and form an integral part of our methodology. Their aim is to hybridize feature sets for classification while exploiting each type of unique merits derived from DenseNet201 or contour or color histogram features. The first of these would use DenseNet201 features exclusively as the base learner. The second would be using those features solely derived from DenseNet201 in conjunction with sundry classifiers. The third would be combined DenseNet201 features and contour features, thus fusing high-level pattern recognition from DenseNet201 with structural information given by the contours. The fourth combines the features of DenseNet201 with the color histogram feature-the deep learning derived patterns with color-based textural information. Clearly, the exploration of these various combinations allows us to gauge which sets of features are the most effective for classifying lung and colon images.

V. MACHINE LEARNING (ML) ALGORITHMS

In the field of lung and colon cancer (LC AND CC) image classification, choosing the right machine learning (ML) models is crucial for achieving optimal performance and reliability. Our methodology integrates a variety of classifiers, each with its own distinct advantages and features. Below is a brief overview of the models selected for this study:



5.1 K-NEAREST NEIGHBORS (KNN)

An intuitive and instance-based learning algorithm, KNN classifies according to the distance of the nearest samples in the training set. It is particularly suited for applications in which decision boundaries are non-linear or irregularly shaped.

5.2LIGHT GRADIENT BOOSTING MACHINE (LGBM)

LGBM is a tree-based model that implements gradient boosting in an efficient manner. LGBM is known for high efficiency with large data sets, especially for the problem of imbalanced data so common in medical image classification

5.3CATBOOST

CatBoost is a gradient-boosting algorithm for both categorical and categorical data, which allows it to be well suited for datasets with mixed blend types. It has a lot of strength and readily allows for handling missing data, making it a good fit for complex medical imaging datasets

5.4EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is a state-of-the-art implementation of the gradient boosting algorithm that is designed to be fast and scalable. The performance of XGBoost in various machine learning competitions has established it as a powerful classification technique granting fine control of model tuning

5.5DECISION TREES (DT)

Decision trees are simplistic yet interpretable models that partition data based on specific criteria at every node. They can provide insight into the classification decision-making processes, helping clinicians understand which features play the most significant role in classifying medical images

5.6RANDOM FOREST (RF)

Random Forest is an ensemble learning method where multiple decision trees are built at training time to combat overfitting and boost the classification accuracy. The model is versatile, applying itself equally well to both linear and non-linear data

5.7MULTINOMIAL NAIVE BAYES (MULTINOMIALNB):

The Multinomial Naive Bayes classifier is geared toward the situation wherein the data follows the multinomial distribution. The classifier assumes that the features are independent, thus simplifying computations. This method performs exceptionally well in tasks with large feature sets and fits very nicely to cases involving high-dimensional medical image data like ours

5.8 SUPPORT VECTOR MACHINE (SVM):

SVM is an advanced classifier that selects the hyperplane with maximum margin separating classes in the feature space. It performs very well for both linear and non-linear data, especially in the high-dimensional data particularly suitable for the intricacies involved in image classification tasks

VI. *EVALUATION METRICS*

The models will be assessed using several evaluation metrics, including accuracy, average specificity, processing time (S), precision, recall, and F1-score, as defined in equations (1) to (5). These metrics offer a well-rounded evaluation of each model's performance, considering aspects such as generalization ability, computational efficiency, and the trade-off between precision and recall.

$$\text{Accuracy} = (1) \quad \frac{TP+TN}{TN+TP+FP+FN}$$

$$\text{Precision} = (2) \quad \frac{TP}{TP+FP}$$

$$\text{Recall} = (3) \quad \frac{TP}{TP+FN}$$

$$\text{F1} = (4)\ 2\times \frac{Recall \times Precision}{Recall+Precision}$$

$$\text{Specificity} = (5)\ \frac{TN}{TN+FP}$$

VII. RESULTS AND ANALYSIS

The research and testing of machine learning algorithms for multi-class classification of lung and colon cancers (LC AND CC) based on the LC AND CC25000 dataset yield important findings. Several metrics such as accuracy, specificity, precision, recall, and F1-score were used to evaluate the models. The evaluation of the models was done based on the three different scenarios of feature extraction (FE): alone with DenseNet201, with contour features as an adjunct to DenseNet201, and with histogram features as an adjunct to DenseNet201.

Algorithm	Accuracy	Avg Specificity	Time (S)	Precision	Recall	F1-score
DenseNet201	0.9733	0.4934	6908.51	0.97 MA	0.97 MA	0.97 MA
				0.97 WA	0.97 WA	0.97 WA

TABLE 1. The performance metrics results of the DenseNet201 base learner on the test set. MA (Micro Average), WA (Weighted average).

IMPACT OF FEATURE COMBINATION METHODS ON MULTI-CLASSIFICATION

The coalescence of the contour features with DenseNet201 has been proving this consistent performance level across the models, KNN further excelling in accuracy and efficiency as shown in Table 1. The performance of SVM drastically reduced with the addition of contour features, whose accuracy was down to 33.1% and shows poor adaptability regarding this feature combination. Models such as LGBM, CatBoost, XGBoost scored high even after additional contour feature integration and were proven to be insensitive to the added contour features. DT and RF again showed decent results where the RF performed better because of its ensemble method compared to DT. MultinomialNB could not perform well under the complexity of the combined features but was still efficient in computational time. The incorporation of contour features with DenseNet201 has resulted in consistent performance across different models, with KNN proving to be very accurate and efficient as shown in Table1. In contrast, there was a marked drop in the performance of SVM under this condition, with accuracy dropping to 33.1% and nothing specified with respect to specificity, as it was an indication of a poor fit to this feature combination. Models such as LGBM, CatBoost, and XGBoost proved to have held up pretty well and were not as sensitive to the inclusion of contour features. Decision Trees (DT) and Random Forest (RF) continued scoring good results, though performance of DT was only moderate: an indication that more complex contour features did not perform well with it. On the other hand, because of the nature of its ensembles, RF was able to record higher accuracy compared to DT. Multinomial Naive Bayes (MultinomialNB) was found not to be able cope with the increased complexity of the combined features resulting in lower performance even though it still proved to be computationally efficient.

Integration of histogram features with DenseNet201 resulted in a significant improvement of the model performance as shown in Table1. In addition, the KNN algorithm achieved an impressive accuracy of 99.68% with perfect specificity. This feature combination improved the accuracy and specificity of the DT and RF models as compared to the classification done using DenseNet201 features alone. The accuracy and specificity of the DT model have improved, so according to this it can be said that the histogram features with the color textural information have proven to be in accordance with the decision making of the model. That's a richer perspective that RF has taken with this integrated approach, demonstrating a superior level of accuracy and robustness over DT. Likewise, adding histogram features has aided LGBM, as well as XGBoost and CatBoost to produce similarly good results. However, as with all features, MultinomialNB has been the least performer in comparison, an indication of its inefficiency on tackling the complexity a typical image classification would require in this scenario. SVM has shown quite impressive performance as result of other histogram features added to it, which in turn proves its ability to effectively distinguish data points in higher dimensions and makes it useful in critical areas where the overhead costs in computing are justified.

The confusion matrices are portraying the performance of all ML solutions at the amalgamation of DenseNet201 features and histogram features. These serve as illustrative tools for drawing analytical comparisons of the modelling and classification performance of each model about its class. This would help in evaluating their predictive powers. It reveals the strengths and weaknesses of the model with regard to the trends on misclassification which would lead to insights into the areas of improvements. Very few instances

were misclassified by KNN: only three benign lung and colon tissues, misclassified as being lung and colon squamous cell carcinomas, whereas almost perfect classification exists for lung and colon adenocarcinomas and lung and colon squamous cell carcinomas. Both CatBoost and XGBoost models were found to be high-performing ones in their errorless classification of lung and colon adenocarcinomas. However, LGBM performed well but had difficulties in distinguishing between benign lung and colon tissues and lung and colon squamous cell carcinomas, differentiated with 15 misclassifications. The significant misclassifications done by Decision Tree and Naive Bayes were mostly confused benign tissues between lung and colon with carcinoma tissues of lungs and colon as well as adenocarcinomas, indicating that the models were weak in deciphering between these classes at finer levels. There were classification errors between RF and SVM concerning benign lung and colon tissues being misclassified as lung and colon squamous cell carcinomas.

They further confuse all models of the ML with histogram features when combined with DenseNet201 features as can be seen from all confusion matrices. Each of the models could also be represented in terms of how they classify and distinguish different classes, which practically help in assessing its predictive power. They have also shown the trend of misclassification in the models and areas for improvement. KNN misclassifies a few instances comparatively, and only 3 benign lung and colon tissues are classified as lung and colon squamous cell carcinomas, while for adenocarcinomas of both lung and colon, squamous cell carcinomas of lung and colon, near perfect classification is maintained with KNN. These previously mentioned models performed exceptionally in the error-free categorization of lung and colon adenocarcinomas: CatBoost and XGBoost. LGBM also performed well but had some difficulty in differentiating among benign lung and colon tissues with lung and colon squamous cell carcinomas, with 15 misclassifications incurred. Decision Tree and Naive Bayes have shown a significant degree of misclassifications regarding benign tissues of lungs and colon as well as lung and colon squamous cell carcinomas or lung and colon adenocarcinomas, thus indicating they are unable to really make fine distinctions between these class levels. Misclassifications occurred on RF and SVM, especially concerning benign tissue that is misclassified to lung and colon squamous cell carcinomas.

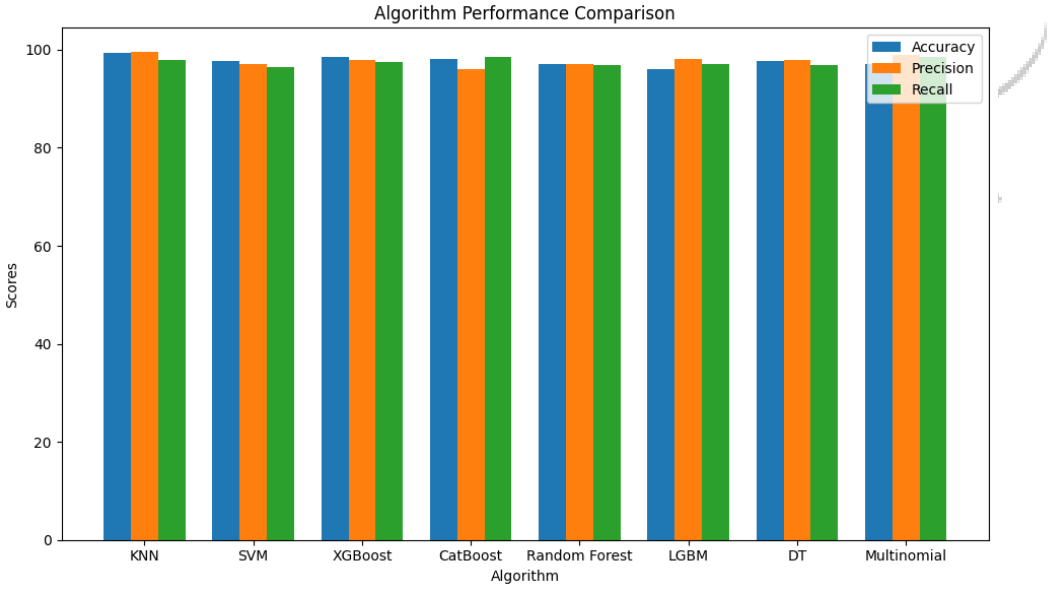


Figure 3 Accuracy and results of DenseNet201features integrated with contour features in conjunction with ML models

Algorithm	Accuracy	Avg Specificity	Time (S)	Precision	Recall	F1-score
LGBM	0.9863333	0.99	104.07	0.99 MA	0.99 MA	0.99 MA
				0.99 WA	0.99 WA	0.99 WA
CatBoost	0.9881666	0.99	1246.11	0.99 MA	0.99 MA	0.99 MA
				0.99 WA	0.99 WA	0.99 WA
XGBoost	0.9866666	0.99	84.26	0.99 MA	0.99 MA	0.99 MA
				0.99 WA	0.99 WA	0.99 WA
KNN	**0.9901666**	0.99	**0.01**	0.99 MA	0.99 MA	0.99 MA
				0.99 WA	0.99 WA	0.99 WA
DT	0.9336666	0.95	30.61	0.93 MA	0.93 MA	0.93 MA
				0.93 WA	0.93 WA	0.93 WA
RF	0.9705	0.98	48.75	0.97 MA	0.97 MA	0.97 MA
				0.97 WA	0.97 WA	0.97 WA
MultiNB	0.9211666	0.92	0.08	0.92 MA	0.92 MA	0.92 MA
				0.92 WA	0.92 WA	0.92 WA
SVM	0.978	0.99	13.51	0.98 MA	0.98 MA	0.98 MA
				0.98 WA	0.98 WA	0.98 WA

TABLE 2. Performance metrics results of DenseNet201 features in conjunction with ML models. MA (micro average), WA (weighted average).

VIII. DISCUSSION

This study investigated recently modernized AI techniques for analysis of histopathological images and their usefulness for the early detection and classification of lung and colon cancer (LC and CC). The results are promising as regards the mortality impacts and the difficulties in early diagnosis associated with LC AND CC. The paper evaluates the performance of the machine-learning algorithms like KNN, CatBoost, XGBoost, and LGBM for this multiclass and binary classification. This emphasizes the need to choose algorithms based on requirements in terms of accuracy, specificity, and efficiency under the required cost of computation.
For instance, KNN boasts both high accuracy and efficiency in multi-class and binary-classification tasks with less misclassification, making it very useful for real-time applications where accuracy and speed are prioritized. Conversely, accuracy in prediction is attained through CatBoost and XGBoost but coupled with long computation for inference making it suitable for usage in applications where precision is paramount, while time is not a significant constraint. Applications that require long-prediction time but pose restrictions on false positives are preferred to be equipped with either CatBoost or XGBoost.

Decision Tree, Naive Bayes, Random Forest, and SVM models performed much poorer than KNN, CatBoost, and XGBoost in terms of misclassification and efficiency as well as being computationally less efficient. This study also shows that integration of DenseNet201 with contour and histogram features is valuable since it improves the power of classifiers by combining different methods of feature extraction (FE). Indeed, this hybrid approach has shown promise in improving the accuracy of LC AND CC diagnostics**.**
Late studies indicate that, besides classification accuracy, one other principal aspect to consider for a real-world application is computational efficiency. CatBoost gives effective accuracy but takes longer to compute, which emphasizes the importance of optimization for real-world scenarios where accuracy and efficiency are equally important. The difficulties in distinguishing the adenocarcinoma and squamous cell carcinoma subtypes in lung and cervical cancer highlight the intricate nature of their classification and allow for further refinement of the approaches and feature extraction methods through investigation.

This study makes a comparison of the commented model with other state-of-the-art models to classify lung and cervical cancer histopathological images, as Table 6 outlines. The model proposed here stands out with an impressive accuracy of 99.68%, surpassing the other models. This improved performance can be attributed to successfully combining cutting-edge feature extraction techniques with DenseNet201 to KNN's powerful classification. This hybrid approach brings not only increased accuracy to the model but also some

computational efficiencies. Compared with AlexNet CNN with Histogram Equalization or DenseNet121 FE with RF, this model improves accuracy, validating the efficacy of the hybrid FE strategy. By combining DenseNet201 with color histogram features, the proposed method magnified the ability to distinguish subtle features in histopathological images that are vital for early detection of lung and cervical cancer.

IX. RESEARCH GAP, LIMITATIONS, AND FUTURE WORK

The study attempts to bridge a considerable gap in contemporaneous academic literature on accuracy and efficiency in the diagnosis of lung and colon cancers (LC AND CC) using histopathological images. While great strides have been made toward diagnostic technologies, there remains a huge unresolved problem in the diagnosis of lung and colon adenocarcinomas, lung and colon squamous cell carcinomas, and non-cancerous lung and colon tissues with minimum human intervention in an accurate and expeditious manner. The research attempts to employ DenseNet201 alongside color histogram methods and a variety of machine learning (ML) techniques to increase the diagnostic accuracy--clearly filling an important void with respect to the automated detection of cancers from medical images.

Nonetheless, the method developed in this work has limitations. While the LC AND CC25000 dataset represents a very large dataset, it may not naturally reflect the variability seen in the wider clinical setting. The second weakness is that the verification results may not generalize, solely relying on this dataset from various image acquisition settings. Such reliance could bias selection because this dataset does not fully encompass the full range of LC AND CC histopathologies typically seen in everyday medical practice. The performance metrics related to the set of algorithms are dataset-specific; additional warrants and investigations have to be initiated to extend the findings on cancers or their subtypes.

Future work to address these limitations should emphasize the need to diversify the dataset concerning a variety of histopathological images. In this way, selection bias may be reduced, and generalizability of the model may be improved. Additionally, the adoption of different data augmentation techniques could help reduce the processing time and thereby facilitate its potential use in real-time clinical applications. Applying better methods for image preprocessing should help mitigate some impact of image quality on the results. Also, by focusing on relevant feature selection strategies, the model performance can be enhanced due to selection of relevant features. Image segmentation may allow for more focused targeting of areas of interest while rendering better detection of subtle signs of cancer. Finally, advanced exploration of feature extraction (FE) methods, which will enhance sensitivity to the identification of pathological characteristics in histopathology images, would also be vital. Therefore, this set of improvements will be relevant toward the enhancement of the diagnostic accuracy and clinical utility of these ML models in LC AND CC detection as they would fulfill the very high demands imposed by clinical practitioners.

Ref	Year	Dataset	Efficacy	Strengths	Drawbacks	Hardware
[11]	2023	LC25000	VGG-19 + Handcrafted features + ANN 99.64% accuracy, 99.85% sensitivity, 100% specificity and precision.	• Hybrid AI systems integrating CNN models with handcrafted features. • VGG-19 + Handcrafted was optimal. • Applied on colon and lung. • High performance metrics. indicating the model's reliability.	• Single dataset, results may not generalize to other datasets. • Time computation not available. • Huge number of features.	Not provided
[12]	2020	LC25000	Pre-trained CNN models with visualization of class activation and saliency maps accuracy of 96-100% in classifying malignant vs benign tumors.	• Application of visualization techniques (GradCAM, SmoothGrad) for interpretability. • Effective use of pre-trained CNN models.	• Single dataset. • Time computation not available. • Binary classification. • Lack of detailed about computational environment.	Not provided
[13]	2022	LC25000	DenseNet121 FE + RF 98.60% accuracy, 98.63% precision, 98.60% recall, an f1-score of 0.986.	• Evaluates classifier performance using multiple metrics. • Applied on colon and lung. • Comprehensive comparison FE.	• Single dataset. • Time computation not available.	Python 3.8, IBM Intel Core i-7–6700 CPU @ 3.40 GHz processor, 8 GB RAM, NVIDIA GeForce GPU.
[16]	2021	LC25000	Unsharp masking for image sharpening: 2D Fourier and wavelet transforms for FE + CNN Model 96.33% accuracy, 96.39 % precision.	• Uses fourier and wavelet transforms to extract complementary feature sets. • Enhance CNN by Employing a custom-designed 4-channel CNN architecture. • Applied on colon and lung. • High performance metrics.	• Single dataset. • Time computation not available. • Computational resources.	Not provided

[17]	2021	LC25000	Capsule network + conventional and separable CNNs 99.58% accuracy, 98.66% precision.	• Allows the model to learn features from both unprocessed and preprocessed images. • Use of capsule networks with convolutional layers. • Improve the overall feature learning process of the model. • Applied on colon and lung.	• Single dataset. • Time computation not available. • Computational complexity due to the dual-input approach.	Windows 10 PC, Nvidia GeForce GTX 1060, 16 GB RAM, Intel Ci7 64-bit, Keras and TensorFlow.
[18]	2021	LC25000	CNN model using triplet loss 99.08% accuracy with DenseNet121.	• Comprehensive exploration of various CNN architectures. • Application of triplet loss improves the differentiation between the classes. • Applied on colon and lung.	• Single dataset. • Time computation not available. • Specific hardware details are not provided affect reproducibility.	Python
[19]	2020	LC25000	CNN model 96.11% training accuracy, 97.20% validation accuracy.	• Improving the quality of input data for the CNN model by image preprocessing.	• Single dataset. • Lack comparison with other algorithms. • Specific hardware details are not provided affect reproducibility.	Google Colaboratory GPU
[20]	2022	LC25000	Initial accuracy 89%, improved to 98.4% by AlexNet CNN +CISP (Histogram Equalization).	• Improves classification using CSIP. • Minimizes computational costs. • Applied on colon and lung. • Computational efficiency.	• Single pretrained model. • Single dataset. • Limited description of CSIP. • Time computation not available.	Not provided
[21]	2020	LC25000	Shallow CNN 97.92% and 96.95% accuracy (lung and colon respectively).	• Integration of DL. • Develop CNN models. • Effective use of shallow CNN architecture.	• Single dataset. • Lacks some implementation details. • Lack of detailed about computational environment.	Google's Colab TensorFlow
[22]	2022	LC25000	Train CNN model and used explainable DL (GradCAM) 97.11% accuracy.	• Highly accurate, explainable. • Applied explainable DL techniques. • Highlighting specific image areas used for classification.	• Single dataset. • Time computation not available.	Google's Colab TensorFlow
[23]	2022	Kaggle	Ensemble learning techniques (XGBoost, LightGBM, bagging, and AdaBoost) 94.42% accuracy.	• Effectiveness of XGBoost in LC prediction. • Comprehensive evaluation of ensemble learning techniques.	• Single dataset and small. • Time computation not available.	Not provided
[24]	2023	LC25000	Multi-level CNN (ML-CNN) training accuracy: 64%, validation accuracy: 89%.	• Handle the heterogeneity in lung nodule sizes and morphologies. • Leveraging multi-scale convolution for improved FE.	• Single dataset and small. • Time computation not available.	Python 3.X and Google Colab. provided a Jupyter notebook - GPU
[25]	2023	LC25000	Grey wolf optimization (GWO) + Invasive Weed optimization (IWO) + hyperparameter tuning RAdam + DT accuracy of 91.57%.	• Integration of PSO and GWO for FE. • The use of hyperparameter tuning methods to improve accuracy.	• Single dataset. • Time computation not available. • Binary classification. • Lack of detailed about computational environment.	Not provided
[26]	2023	LC25000	Histogram of oriented gradient (HOG) + hyperparameter tuning green anaconda optimization (GAO) + improved graph neural network (IGNN) accuracy of 98.9%.	• Employed gabor filter for preprocessing and MEM for segmentation. • Introduced a novel IGNN model optimized by GAO.	• Single dataset. • Lack of detailed about computational environment.	Not provided

TABLE 3. Comprehensive comparative Meta analysis of efficacy, strengths, and drawbacks among diverse imaging techniques applied to the diagnosis and treatment of Lung and colon cancer (LC AND CC).

X. CONCLUSION

This research intends to improve the early detection and classification of lung and colon cancer (LC AND CC) using sophisticated artificial intelligence (AI) methods. This study integrates DenseNet201 for deep feature extraction (FE) with color histogram features, while also analyzing the data through multiple machine learning (ML) algorithms, particularly KNN, which delivered excellent performance. The model applied to the LC AND CC25000 dataset managed an extraordinary accuracy of 99.68%, far better than other existing models. The high accuracy of this model becomes important due to the very high mortality rate of LC AND CC and the difficulties in early diagnosis. The study also highlights the importance of selecting the algorithm best suited for the unique needs of the task-KNN for real-time applications due to speed and accuracy; CatBoost for problems where accuracy is essential, even with longer running time.
The synergy of different feature extraction techniques not only promotes class discrimination via classifier performance but also flexibility to adapt to different types of cancer, as seen from its application to the BreakHis dataset on breast cancer histopathology, suggesting that this avenue could be pursued in other subspecialties of oncology. The study further presents considerations about the trade-off between operational speed and classification accuracy inherent in LC AND CC subtype differentiation, thus calling for continued enhancement of algorithms and feature extraction techniques.

Introducing a novel method to LC AND CC diagnosis, this study becomes instrumental in paving a path for AI-led initiatives in oncology diagnostics. It sparks a thrilling line of inquiry and initiative toward ensuring more accurate and effective cancer diagnosis and classification. As the endeavor matures, the synergy among data scientists, medical professionals, and oncologists will remain imperative in converting these technological breakthroughs into clinically meaningful advances.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, ‘‘Cancer statistics, 2023,’’ *CA, Cancer J. Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[2] Amer. Cancer Soc. (Oct. 18, 2023). *Cancer Statistics Center, American Cancer Society*. Accessed: Dec. 5, 2023. [Online]. Available: https://cancerstatisticscenter.cancer.org/#!/

[3] F. Venuta, D. Diso, I. Onorati, M. Anile, S. Mantovani, and E. A. Rendina, ‘‘Lung cancer in elderly patients,’’ *J. Thoracic Disease*, vol. 8, no. S11, pp. S908–S914, Nov. 2016, doi: 10.21037/jtd.2016.05.20.

[4] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, ‘‘The global burden of lung cancer: Current status and future trends,’’ *Nature Rev. Clin. Oncol.*, vol. 20, no. 9, pp. 624–639, Sep. 2023, doi: 10.1038/s41571-023-00798-3.

[5] (Jun. 26, 2023). *World Health Organization: WHO*. Accessed: Jan. 4, 2024. [Online]. Available: https://www.who.int/news-room/factsheets/detail/lung-cancer

[6] Lung Cancer Statistics | CDC. *Centers for Disease Control and Prevention*. Accessed: Feb. 4, 2024. [Online]. Available: https://www. cdc.gov/cancer/lung/statistics/index.htm

[7] SEER. (2023). *Common Cancer Sites—Cancer Stat Facts*. Accessed: Dec. 20, 2023. [Online]. Available: https://seer.cancer.gov/ statfacts/html/common.html

[8] A. J. Tybjerg, S. Friis, K. Brown, M. C. Nilbert, L. Morch, and B. Køster, ‘‘Updated fraction of cancer attributable to lifestyle and environmental factors in Denmark in 2018,’’ *Sci. Rep.*, vol. 12, no. 1, p. 549, Jan. 2022, doi: 10.1038/s41598-021-04564-2.

[9] M. Del Re, E. Rofi, G. Restante, S. Crucitta, E. Arrigoni, S. Fogli, M. Di Maio, I. Petrini, and R. Danesi, ‘‘Implications of Kras mutations in acquired resistance to treatment in NSCLC,’’ *Oncotarget*, vol. 9, no. 5, pp. 6630–6643, Jan. 2018, doi: 10.18632/oncotarget.23553.

[10] (Mar. 22, 2022). *Lung Cancer—Diagnosis and Treatment, Mayo Clinic*. Accessed: Jan. 11, 2024. [Online]. Available: https://www. mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627

[11] M. Al-Jabbar, M. Alshahrani, E. M. Senan, and I. A. Ahmed, ‘‘Histopathological analysis for detecting lung and colon cancer malignancies using hybrid systems with fused features,’’ *Bioengineering*, vol. 10, no. 3, p. 383, Mar. 2023, doi: 10.3390/bioengineering10030383.

[12] S. Garg and S. Garg, ‘‘Prediction of lung and colon cancer through analysis of histopathological images by utilizing pre-trained CNN models with visualization of class activation and saliency maps,’’ in *Proc. 3rd Artif. Intell. Cloud Comput. Conf.*, New York, NY, USA, Dec. 2020, pp. 38–45, doi: 10.1145/3442536.3442543.

[13] 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, May 2022, Art. no. 103596, doi: 10.1016/j.bspc.2022.103596.

[14] (Jun. 19, 2019). *How Do Cancer Cells Grow and Spread?* Accessed: Dec. 5, 2023. [Online]. Available: https://ncbi.nlm.nih.gov/books/NBK279410/

[15] R. W. Pettit, J. Byun, Y. Han, Q. T. Ostrom, J. Edelson, K. M. Walsh, M. L. Bondy, R. J. Hung, J. D. McKay, and C. I. Amos, ‘‘The shared genetic architecture between epidemiological and behavioral traits with lung cancer,’’ *Sci. Rep.*, vol. 11, no. 1, p. 17559, Sep. 2021, doi: 10.1038/s41598-021-96685-x.

[16] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, ‘‘A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,’’ *Sensors*, vol. 21, no. 3, p. 748, Jan. 2021, doi: 10.3390/s21030748.

[17] M. Ali and R. Ali, ‘‘Multi-input dual-stream capsule network for improved lung and colon cancer classification,’’ *Diagnostics*, vol. 11, no. 8, p. 1485, Aug. 2021, doi: 10.3390/diagnostics11081485.

[18] N. Baranwal, P. Doravari, and R. Kachhoria, ‘‘Classification of histopathology images of lung cancer using convolutional neural network (CNN),’’ in *Disruptive Developments in Biomedical Applications (CNN)*. Boca Raton, FL, USA: CRC Press, 2022, pp. 75–89.

[19] B. K. Hatuwal and H. C. Thapa, ‘‘Lung cancer detection using convolutional neural network on histopathological images,’’ *Int. J. Comput. Trends Technol.*, vol. 68, no. 10, pp. 21–24, Oct. 2020, doi: 10.14445/22312803/ijctt-v68i10p104.

[20] S. Mehmood, T. M. Ghazal, M. A. Khan, M. Zubair, M. T. Naseem, T. Faiz, and M. Ahmad, ‘‘Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing,’’ *IEEE Access*, vol. 10, pp. 25657–25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[21] S. Mangal, A. Chaurasia, and A. Khajanchi, ‘‘Convolution neural networks for diagnosing colon and lung cancer histopathological images,’’ 2020, *arXiv:2009.03878*.

[22] J. Civit-Masot, A. Bañuls-Beaterio, M. Domínguez-Morales, M. Rivas-Pérez, L. Muñoz-Saavedra, and J. M. Rodríguez Corral, ‘‘Non-small cell lung cancer diagnosis aid with histopathological images using explainable deep learning techniques,’’ *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107108, doi: 10.1016/j.cmpb.2022.107108.

[23] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, ‘‘Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,’’ in *Proc. IEEE World AI IoT Congr.*, Jul. 2022, pp. 187–193, doi: 10.1109/AIIoT54504.2022.9817326.

[24] M. Ramesh, S. Maheswaran, S. Theivanayaki, K. Kodeeswari, L. Krishnasamy, and N. Sriram, ‘‘Efficient lung cancer classification on multi level convolution neural network using histopathological images,’’ in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2023, pp. 1–7, doi: 10.1109/icccnt56998.2023.10307852.

[25] K. Shanmugam and H. Rajaguru, ‘‘Exploration and enhancement of classifiers in the detection of lung cancer from histopathological images,’’ *Diagnostics*, vol. 13, no. 20, p. 3289, Oct. 2023, doi: 10.3390/diagnostics13203289.

[26] S. Dinesh Krishnan, D. Pelusi, A. Daniel, V. Suresh, and B. Balusamy, ‘‘Improved graph neural network-based green anaconda optimization for segmenting and classifying the lung cancer,’’ *Math. Biosc. Eng.*, vol. 20, no. 9, pp. 17138–17157, Sep. 2023, doi: 10.3934/mbe.2023764.

[27] S. Chakraborty and K. Mali, ‘‘An overview of biomedical image analysis from the deep learning perspective,’’ in *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*. Hershey, PA, USA: IGI Global, 2023, ch. 3, pp. 43–59, doi: 10.4018/978-1-6684-75447.ch003.

[28] A. A. Borkowski, M. M. Bui, L. Brannon Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, ‘‘Lung and colon cancer histopathological image dataset (LC25000),’’ 2019, *arXiv:1912.12142*.

[29] R. C. Gonzalez, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.

[30] D. Singh and B. Singh, ‘‘Investigating the impact of data normalization on classification performance,’’ *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 105524, doi: 10.1016/j.asoc.2019.105524.

[31] A. Güneş, H. Kalkan, and E. Durmuş, ‘‘Optimizing the color-to-grayscale conversion for image classification,’’ *Signal,ImageVideoProcess.*, vol. 10, no. 5, pp. 853–860, Jul. 2016.

[32] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Stamford, CT, USA: Cengage Learning, 2014.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ‘‘Densely connected convolutional networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[34] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.